# Applied Regression Analysis

## Week 4

1. Homework week 3: highlights
2. Polynomial regression I
3. Polynomial regression II
4. Polynomial regression III
5. Example: Dose-response study / assessing multicollinearity
6. Example: Potential energy
7. Homework

Stanley Lemeshow, Professor of Biostatistics

*College of Public Health, The Ohio State University*

The Ohio State University

# WEEK 4: POLYNOMIAL REGRESSION

A polynomial of order $k$ in $x$ is an expression of the form

$$y = c_0 + c_1 x + c_2 x^2 + c_3 x^3 + \cdots + c_k x^k$$

where the $c$'s and $k$ are constants.

When $k = 1$ we had

$$y = c_0 + c_1 x \longleftarrow \text{straight line}$$

Let us now focus on the 2nd order polynomial $(k = 2)$

$$y = c_0 + c_1 x + c_2 x^2$$

These are mathematical models.

The statistical model for the $k$ = 2 case can be expressed in one of two ways:

mean of $y$ at a given $x$

$$\mu_{y|x} = \beta_0 + \beta_1 x + \beta_2 x^2$$

or

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

unknown parameters (regression coefficients)

Error component

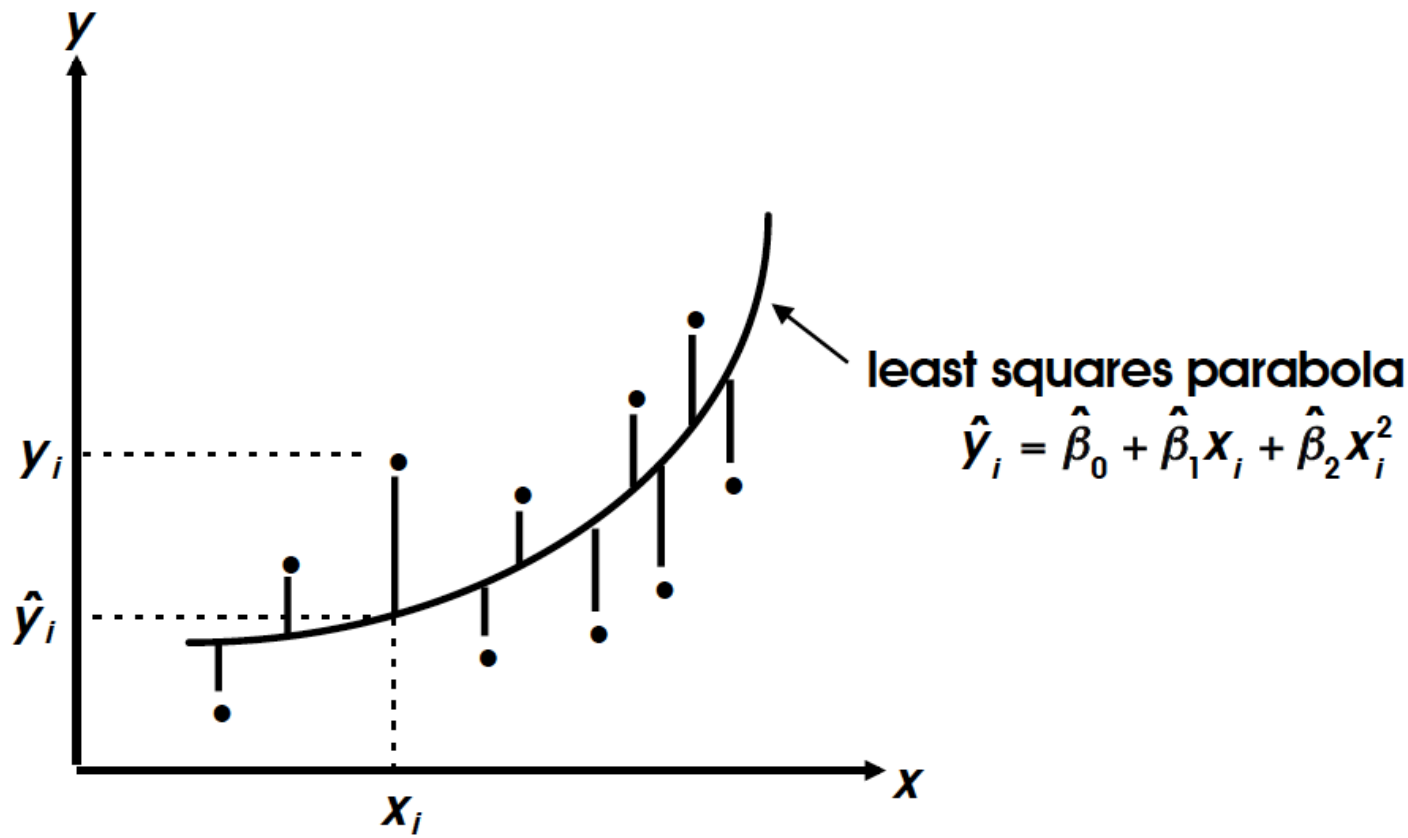
Let us now use the method of least squares to obtain estimates for the regression coefficients in the parabolic (2nd degree) model.

The estimated parabola may be written as

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$$

and

$$SSE = \sum_{i=1}^{n} \left( y_i - \hat{y}_i \right)^2 = \sum_{i=1}^{n} \left( y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i - \hat{\beta}_2 x_i^2 \right)^2$$




$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ are chosen so that the SSE is smaller than for any other choice of $\beta$'s.

Instead of presenting here the precise formulas for $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ (which can get very complex-particularly when $k$ is large) it is assumed that you will be doing this work via computer.

For the age-SBP example with the outlier removed ($n$ = 29) we obtain from the computer:

$$\hat{\beta}_0 = 113.41$$

$$\hat{\beta}_1 = 0.088$$

$$\hat{\beta}_2 = 0.010$$

Hence, the fitted model is

$$\hat{y} = 113.41 + 0.088x + 0.010x^2$$

Recall that for these $n$ = 29 individuals, the straight-line model was

$$\hat{y} = 97.08 + 0.95x$$

Now, the essential results based on fitting a 2nd - (or higher) order polynomial model can be summarized in an ANOVA table.

As was true for the 1st order polynomial model,

$$\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2 + \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$$

$$\mathrm{SSY} = \left(\mathrm{SSY} - \mathrm{SSE}\right) + \mathrm{SSE}$$

Total SS = Due regression SS + residual SS

Then the ANOVA Table is

| Source | df | SS | MS | F |
|---|---|---|---|---|
| Due Regression | $k = 2$ | $SSY - SSE = 6273.40$ | $\frac{SSY - SSE}{k} = 3136.70$ | 35.37 |
| Due Residual | $n - k - 1 = 26$ | $SSE = 2306.05$ | $\frac{SSE}{n-1-k} = 88.69$ | $(p < .001)$ |
| Total $r^2 = .731$ | $n - 1 = 28$ | $SSY = 8579.45$ | | |

Now recall that in the straight-line model with the outlier removed we had

| Source | df | SS | MS | F |
|---|---|---|---|---|
| Due Regression | 1 | 6110.10 | 6110.10 | 66.81 |
| Due Residual | 27 | 2469.35 | 91.46 | $(p < .0001)$ |
| Total $r^2 = .712$ | 28 | 8579.45 | | |

These tables give rise to the following for the 2nd order polynomial model:

| Source | | df | SS | MS | F |
| --- | --- | --- | --- | --- | --- |
| Regression | $x$ | 1 | 6110.10 | 6110.10 | 66.81 $= \frac{6110.1}{91.46}$ |
| | $x^2 \mid x$ | 1 | 163.30 | 163.3 | 1.84 $= \frac{163.30}{88.69}$ |
| Residual | | 26 | 2306.05 | 88.69 | |
| Total | | 28 | 8979.45 | | |

computed by subtraction

note: as usual, the residual sum of squares SSE is divided by its degrees of freedom to yield an estimate of $\sigma^2$

i.e.,

$$\text{MS residual} = s^2_{y|x} = \frac{1}{n-3}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$$

follows general rule = *n* - # estimated regression coefficients

There are 2 basic inferential questions associated with 2nd order polynomial regression:

(1) Is the overall regression significant?

(2) Does the 2nd order model explain significantly more than that achieved by the straight-line model?

## (1) Test for overall regression

$H_0$ : There is no significant overall regression using $x$ and $x^2$

$H_a$ : There is a significant overall regression

we use $F = \dfrac{\text{MS regression}}{\text{MS residual}}$

and compare this to the $F(2, n-1-k)$

In our example

$$F = 35.37 \text{ and } F_{.999}(2,26) = 9.12 \quad \therefore \text{ reject } H_0 \, (p < .001)$$

This $F$ - test is not equivalent to any $t$ - test

$\left[\text{This is true since } F(1,\upsilon) = t^2(\upsilon) \text{ but } F(2,\upsilon) \neq t\right]$

We can compute the multiple $R^2$

$R^2$ = "squared multiple correlation coefficient"

= proportionate reduction in the error sum of squares obtained using $x$ and $x^2$ instead of the naive predictor $\bar{y}$.

$$R^2 = \frac{\text{SSY} - \text{SSE}\left(2^{\text{nd}} \text{ order model}\right)}{\text{SSY}} = \frac{\text{Due Reg. SS}}{\text{Total SS}}$$

In our example $R^2 = 0.731$. The $F$-test also tests

$$H_0 : R^2 = 0$$
$$\text{vs } H_a : R^2 > 0.$$

As was true for the straight-line model, this one is significant.

## (2) Test for the Addition of $x^2$ Into the Model

$H_0$: The addition of the $x^2$ term to the model does not significantly improve the prediction of $y$ over and above that achieved by the straight-line model.

$H_a$: it does add to the prediction of $y$

note:

$r^2 = .712$ for the straight-line model

$R^2 = .731$ for the second order model

*more variation will always be explained by adding extra terms to the model.

The question here is whether the increase

$$= (.731 - .712) = .019$$

represents a <u>significant</u> increase in the variation explained by the additional term.

(i.e., is .019 enough of an increase to warrant adding the $x^2$ term to the model).

To answer this we compute the extra sum of squares due to the addition of $x^2$. This appeared in the ANOVA table under the source heading "Regression $x^2 | x$".

$$\text{Extra SS due to adding } x^2 = \underset{(2^{nd}\text{ order model})}{\text{SS regression}} - \underset{(1^{st}\text{ order model})}{\text{SS regression}}$$

In our example,

$$\text{SS regression }\left(\text{straight-line model}\right) = 6110.10$$

$$\text{SS regression }\left(2^{\text{nd}}\text{ order model}\right) = 6273.40$$

$$\text{Extra SS due to adding } x^2 \text{ term} = 6273.40 - 6110.10 = 163.30$$

To test $H_0$, we use

$$F = \frac{\left(\text{Extra } SS \text{ due to adding } x^2\right)/1}{MS \text{ residual for } 2^{\text{nd}} \text{ order model}}$$

and this $F$ is compared to the $F\left(1, n-1-k\right)$

## In our example

$$F = \frac{163.30}{88.69} = 1.84$$

and $F_{.90}\left(1,26\right) = 2.91$

in fact $.10 < p < .25$

## Another way to perform this test is to compute

$$t = \frac{\hat{\beta}_2}{\widehat{\text{SE}}\left(\hat{\beta}_2\right)}$$

obtain from computer output

and compare this to a $t\left(n - 1 - k\right)$

```
. use ":Macintosh HD:Desktop Folder:notes1.dta"

. drop if sbp==220
(1 observation deleted)

. regress sbp age

  Source |       SS       df       MS                  Number of obs =      29
---------+------------------------------               F(  1,    27) =   66.81
   Model |  6110.10173     1  6110.10173               Prob > F      =  0.0000
Residual |  2469.34654    27  91.4572794               R-squared     =  0.7122
---------+------------------------------               Adj R-squared =  0.7015
   Total |  8579.44828    28  306.408867               Root MSE      =  9.5633

------------------------------------------------------------------------------
     sbp |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     age |   .9493225   .1161445      8.174   0.000       .7110137    1.187631
   _cons |   97.07708   5.527552     17.562   0.000       85.73549    108.4187
------------------------------------------------------------------------------

. vif

Variable |      VIF       1/VIF
---------+----------------------
     age |      1.00    1.000000
---------+----------------------
Mean VIF |      1.00
```

```
. gen agesq=age*age

. regress sbp age agesq

  Source |       SS       df       MS              Number of obs =      29
---------+------------------------------           F(  2,    26) =   35.37
   Model |  6273.40168     2  3136.70084           Prob > F      =  0.0000
Residual |  2306.0466     26  88.6940999           R-squared     =  0.7312
---------+------------------------------           Adj R-squared =  0.7105
   Total |  8579.44828    28  306.408867           Root MSE      =  9.4178

------------------------------------------------------------------------------
     sbp |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     age |   .0875433   .6453289      0.136   0.893      -1.238949    1.414036
   agesq |   .0099368   .0073232      1.357   0.186      -.0051163    .0249899
   _cons |   113.4097   13.21041      8.585   0.000       86.25533    140.5641
------------------------------------------------------------------------------

. vif

Variable |      VIF       1/VIF
---------+----------------------
     age |      31.83    0.031413
   agesq |      31.83    0.031413
---------+----------------------
Mean VIF |      31.83
```

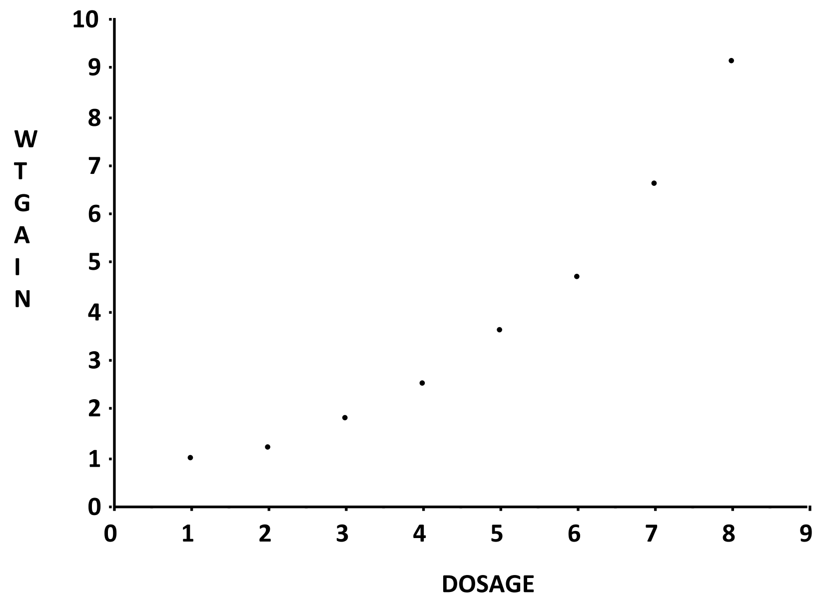
## Another Example

Let $x$ = dose of a certain drug
$y$ = weight gain (in decagrams) after 2 weeks

$n$ = 8 laboratory animals were used and each assigned to one of eight dosage levels of the drug.

| $x$ (Dosage) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| $y$ (Weight Gain) | 1 | 1.2 | 1.8 | 2.5 | 3.6 | 4.7 | 6.6 | 9.1 |

Scatter diagram:



If we had fit a straight-line regression to these data we would find

$$\hat{y} = -1.20 + 1.11x$$

## ANOVA (straight-line model)

| Source | df | SS | MS | F |
|---|---|---|---|---|
| Regression ($x$) | 1 | 52.04 | 52.04 | 61.95 |
| Residual | 6 | 5.03 | 0.84 | |
| Total | 7 | 57.07 | | |

note that $r^2 = 0.912$

and $F = 61.95$ is compared to $F_{.999}(1,6) = 35.51$

i.e., $p < .001$

Let us decide whether or not the addition of the x $^2$ term significantly improves the prediction of *y* over and above that achieved via a straight-line.

$$2^{nd} \text{ order equation: } \hat{y} = 1.35 - 0.41x + 0.17x^2$$

**ANOVA**

| Source | | df | SS | MS | F | |
|---|---|---|---|---|---|---|
| Regression | $x$ | 1 | 52.04 | 52.04 | 61.95 | $\leftarrow F(1,6)$ |
| | $x^2 \mid x$ | 1 | 4.83 | 4.83 | 120.75 | $\leftarrow F(1,5)$ |
| Residual | | 5 | 0.2 | 0.04 | | |
| Total | | 7 | 57.07 | | | |

here $R^2 = .997$

We would like to know whether the increase of $(.997 - .912) = .085$ in $R^2$ represents a significant improvement in the fit.

The test for this is

$$F = \frac{\left(\text{extra } SS \text{ due to adding } x^2\right)/1}{MS \text{ residual for } 2^{\text{nd}} \text{ order model}} = \frac{4.83}{0.04} = 120.75$$

$$\text{and } F_{.999}(1,5) = 47.18 \qquad (p < .001)$$

$$\therefore \text{ reject } H_0$$

Hence, the addition of the $x^2$ term to the model significantly improves the prediction.

Also, the test of the overall 2nd order model is highly significant.

$$F = \frac{\text{MS regression}\left(2^{nd}\text{- order model}\right)}{\text{MS residual }\left(2^{nd}\text{- order model}\right)} = \frac{\left(52.04 + 4.83\right)/2}{0.04}$$

$$= 710.88$$

Hence, the straight-line model is not as good as the 2nd order model.

Can the 2nd order model be improved upon?

- let us add the $x^3$ term to the model and see if it improves the prediction.

## ANOVA (3rd order model)

| Source | | df | SS | MS | F |
|---|---|---|---|---|---|
| Regression | $x$ | 1 | 52.040 | 52.04 | |
| | $x^2 \mid x$ | 1 | 4.830 | 4.83 | |
| | $x^3 \mid x, x^2$ | 1 | 0.140 | 0.14 | 10.00 |
| Residual | | 4 | 0.056 | 0.014 | |
| Total | | 7 | 57.066 | | |

Here $R^2 = .999$

is the increase in $R^2 = (.999 - .997 = .002)$ significant?

$H_0$: the addition of the $x^3$ term is not worthwhile

$$F = \frac{\left(\text{extra SS due to adding } x^3\right)/1}{\text{MS residual for } 3^{\text{rd}} \text{ order model}} = \frac{0.14}{.014} = 10.0$$

and $F \sim F(1,4)$

$$F_{.95}(1,4) = 7.71$$
$$F_{.975}(1,4) = 12.22$$

$.025 < p < .05$

I still wouldn't add $x^3$ since

$(1)$ $R^2$ for the $2^{nd}$ order model was very high = .997

$(2)$ Increse in $R^2$ was only .002

$(3)$ Tolerance suggests multicolinearity

$(4)$ Scatter diagram suggests $2^{nd}$ order model

$(5)$ When in doubt use the simplest model

- this promotes ease of interpretation

Hence the best fitting model is

$$\hat{y} = 1.35 - 0.41x + 0.17x^2$$

with $R^2 = 0.997$

Finally, the computer programs give us the standard errors associated with each $\beta$.

| Coeff $\hat{\beta}_i$ | $s_{\hat{\beta}_i}$ |
|---|---|
| $\hat{\beta}_1 = -.41$ | $s_{\hat{\beta}_1} = .141$ |
| $\hat{\beta}_2 = .17$ | $s_{\hat{\beta}_2} = .015$ |

Using these we can compute confidence intervals

$$\hat{\beta}_i - \left[t_{.975}\left(n-1-k\right)\right]s_{\hat{\beta}_i} \le \beta_i \le \hat{\beta}_i + \left[t_{.975}\left(n-1-k\right)\right]s_{\hat{\beta}_i}$$

95% confidence interval

e.g.,

$$0.17 - \left(\underset{t_{.975}(5)}{2.571}\right)(.015) \le \beta_2 \le 0.17 + (2.571)(.015)$$

$$.13 \le \beta_2 \le .21$$

note that 0 is not in the interval

*t* -tests can also be constructed in the obvious way

```
. regress wtgain dose

  Source |       SS       df       MS              Number of obs =       8
---------+------------------------------           F(  1,      6) =   62.05
   Model |   52.037204     1   52.037204           Prob > F      =  0.0002
Residual |  5.03154917     6  .838591529           R-squared     =  0.9118
---------+------------------------------           Adj R-squared =  0.8971
   Total |  57.0687531     7  8.15267902           Root MSE      =  .91575

------------------------------------------------------------------------------
  wtgain |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    dose |   1.113095   .1413027      7.877   0.000       .7673399    1.458851
   _cons |  -1.196429   .7135439     -1.677   0.145      -2.942408    .5495503
------------------------------------------------------------------------------


. vif

Variable |      VIF       1/VIF
---------+----------------------
    dose |      1.00    1.000000
---------+----------------------
Mean VIF |      1.00
```

```
. regress wtgain dose dosesq

  Source |       SS       df       MS              Number of obs =       8
---------+------------------------------           F(  2,     5) =  722.73
   Model |  56.8720267     2  28.4360133           Prob > F      =  0.0000
Residual |  .196726451     5   .03934529           R-squared     =  0.9966
---------+------------------------------           Adj R-squared =  0.9952
   Total |  57.0687531     7  8.15267902           Root MSE      =  .19836

------------------------------------------------------------------------------
  wtgain |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    dose |  -.4136907   .1410916     -2.932   0.033      -.7763782   -.0510031
  dosesq |   .1696429   .0153035     11.085   0.000       .1303039    .2089819
   _cons |   1.348215    .276736      4.872   0.005       .6368421    2.059587
------------------------------------------------------------------------------

. vif

Variable |      VIF       1/VIF
---------+----------------------
    dose |     21.25    0.047059
  dosesq |     21.25    0.047059
---------+----------------------
Mean VIF |     21.25
```

```
. regress wtgain dose dosesq dosecube

  Source |       SS       df       MS              Number of obs =       8
---------+------------------------------           F(  3,      4) = 1362.82
   Model |  57.0129739     3  19.0043246           Prob > F      =  0.0000
Residual |  .055779265     4  .013944816           R-squared     =  0.9990
---------+------------------------------           Adj R-squared =  0.9983
   Total |  57.0687531     7  8.15267902           Root MSE      =  .11809

------------------------------------------------------------------------------
  wtgain |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    dose |    .379618   .2632868      1.442   0.223      -.3513834    1.110619
  dosesq |  -.0383118   .0660419     -0.580   0.593      -.2216734    .1450498
dosecube |   .0154041   .0048452      3.179   0.034       .0019516    .0288565
   _cons |    .585714   .2909724      2.013   0.114       -.222155    1.393583
------------------------------------------------------------------------------


. vif

Variable |      VIF       1/VIF
---------+----------------------
  dosesq |   1116.59    0.000896
dosecube |    399.01    0.002506
    dose |    208.78    0.004790
---------+----------------------
Mean VIF |    574.79
```