STAT 401A - Statistical Methods for Research Workers

Multiple regression models

Jarad Niemi (Dr. J)

Iowa State University

last updated: November 3, 2014

Multiple regression

Recall the simple linear regression model is

$$Y_i \stackrel{ind}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2)$$

The multiple regression model is

$$Y_i \stackrel{ind}{\sim} N(\beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p}, \sigma^2)$$

where

- Y_i is the response for observation i and
- $X_{i,p}$ is the p^{th} explanatory variable for observation i.

We may also write

$$Y_i \stackrel{ind}{\sim} N(\mu_i, \sigma^2) \quad \text{or} \quad Y_i = \mu_i + e_i, e_i \stackrel{iid}{\sim} N(0, \sigma^2)$$

where

$$\mu_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p}.$$

Explanatory variables

There is a lot of flexibility in the mean

$$\mu_i = E[Y_i|X_{i,1}, \ldots, X_{i,p}] = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p}$$

as there are many possibilities for the explanatory variables $X_{i,1}, \ldots, X_{i,p}$:

- Higher order terms (X^2)
- Additional explanatory variables (X_1 and X_2)
- Dummy/indicator variables for categorical variables ($X_1 = \mathrm{I}()$)
- Interactions ($X_1 X_2$)
    - Continuous-continuous
    - Continuous-categorical
    - Categorical-categorical

Interpretation

Model:

$$Y_i \stackrel{ind}{\sim} N(\beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p}, \sigma^2)$$

The interpretation is

- β_0 is the expected value of the response Y_i when all explanatory variables are zero.
- β_p, $p \neq 0$ is the expected increase in the response for a one-unit increase in the p^{th} explanatory variable when all other explanatory variables are held constant.
- R^2 is the proportion of the variance in the response explained by the model

Higher order terms (X^2)

Let

- Y_i be the distance for the i^{th} run of the experiment and
- H_i be the height for the i^{th} run of the experiment.

Simple linear regression assumes

$$Y_i \stackrel{ind}{\sim} N(\beta_0 + \beta_1 H_i \qquad\qquad\qquad\qquad , \sigma^2)$$

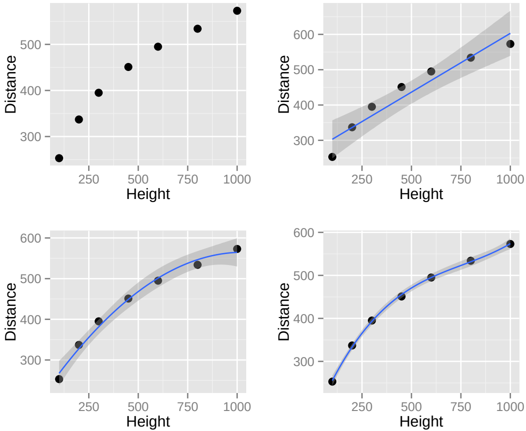
The quadratic multiple regression assumes

$$Y_i \stackrel{ind}{\sim} N(\beta_0 + \beta_1 H_i + \beta_2 H_i^2 \qquad\qquad , \sigma^2)$$

The cubic multiple regression assumes

$$Y_i \stackrel{ind}{\sim} N(\beta_0 + \beta_1 H_i + \beta_2 H_i^2 + \beta_3 H_i^3, \sigma^2)$$

Case1001



SAS code and output

```
DATA case1001;
  INFILE 'case1001.csv' DSD FIRSTOBS=2;
  INPUT distance height;
  height2 = height*height;
  height3 = height*height2;

# PROC REG allows multiple MODEL statements
PROC REG DATA=case1001;
  MODEL distance = height;
  MODEL distance = height height2;
  MODEL distance = height height2 height3;
  RUN;
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	269.71246	24.31239	11.09	0.0001
height	1	0.33334	0.04203	7.93	0.0005
Intercept	1	199.91282	16.75945	11.93	0.0003
height	1	0.70832	0.07482	9.47	0.0007
height2	1	-0.00034369	0.00006678	-5.15	0.0068
Intercept	1	155.77551	8.32579	18.71	0.0003
height	1	1.11530	0.06567	16.98	0.0004
height2	1	-0.00124	0.00013842	-8.99	0.0029
height3	1	5.477104E-7	8.327329E-8	6.58	0.0072

SAS code and output

```
DATA case1001;
  INFILE 'case1001.csv' DSD FIRSTOBS=2;
  INPUT distance height;
  height2 = height ** 2;
  height3 = height ** 3;

PROC GLM DATA=case1001;
  MODEL distance = height height2 height3;
```

```
/* PROC GLM allows the variable construction within the MODEL statement
   and provides nicer output (not shown here) */
DATA case1001;
  INFILE 'case1001.csv' DSD FIRSTOBS=2;
  INPUT distance height;

/* This shorthand puts in H, H^2, and H^3 */
PROC GLM DATA=case1001;
  MODEL distance = height|height|height;

/* This only puts H^3 */
PROC GLM DATA=case1001;
  MODEL distance = height*height*height;
```

R code and output

```
# Construct the variables by hand
case1001$Height2 = case1001$Height^2
case1001$Height3 = case1001$Height^3

m1 = lm(Distance~Height,                case1001)
m2 = lm(Distance~Height+Height2,        case1001)
m3 = lm(Distance~Height+Height2+Height3, case1001)

coefficients(m1)


(Intercept)      Height
   269.7125      0.3333


coefficients(m2)


(Intercept)       Height     Height2
  1.999e+02   7.083e-01  -3.437e-04


coefficients(m3)


(Intercept)       Height     Height2     Height3
  1.558e+02   1.115e+00  -1.245e-03   5.477e-07
```

R code and output

```
# Let R construct the variables for you
m = lm(Distance~poly(Height,3), case1001)
summary(m)



Call:
lm(formula = Distance ~ poly(Height, 3), data = case1001)

Residuals:
      1       2       3       4       5       6       7 
-2.4036  3.5809  1.8917 -4.4688 -0.0804  2.3216 -0.8414 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)        434.00       1.52  286.31  9.4e-08 ***
poly(Height, 3)1   267.12       4.01   66.60  7.5e-06 ***
poly(Height, 3)2   -70.19       4.01  -17.50  0.00041 ***
poly(Height, 3)3    26.38       4.01    6.58  0.00715 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.01 on 3 degrees of freedom
Multiple R-squared:  0.999, Adjusted R-squared:  0.999 
F-statistic: 1.6e+03 on 3 and 3 DF,  p-value: 2.66e-05
```

Longnose Dace Abundance

From `http://udel.edu/~mcdonald/statmultreg.html`:

> *I extracted some data from the Maryland Biological Stream Survey. ... The dependent variable is the number of Longnose Dace (Rhinichthys cataractae) per 75-meter section of [a] stream. The independent variables are the area (in acres) drained by the stream; the dissolved oxygen (in mg/liter); the maximum depth (in cm) of the 75-meter segment of stream; nitrate concentration (mg/liter); sulfate concentration (mg/liter); and the water temperature on the sampling date (in degrees C).*

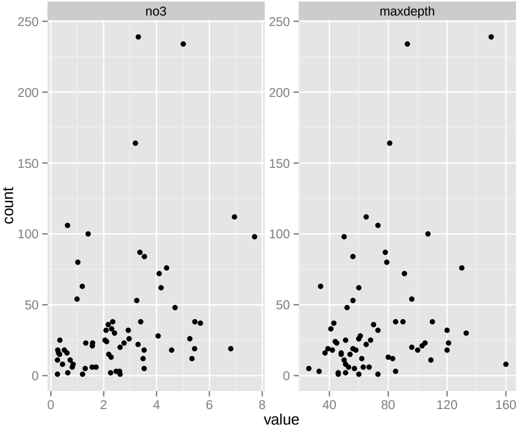
Consider the model

$$Y_i \stackrel{ind}{\sim} N(\beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2}, \sigma^2)$$

where

- Y_i: count of Longnose Dace in stream i
- $X_{i,1}$: maximum depth (in cm) of stream i
- $X_{i,2}$: nitrate concentration (mg/liter) of stream i

Exploratory



```
DATA dace;
  INFILE 'Longnose Dace.csv' DSD FIRSTOBS=2;
  INPUT stream $ count acreage do2 maxdepth no3 so4 temp;

PROC REG DATA=dace;
  MODEL count = maxdepth no3;
  RUN;
```

```
                              The REG Procedure
                               Model: MODEL1
                          Dependent Variable: count

                   Number of Observations Read          67
                   Number of Observations Used          67


                            Analysis of Variance
                                    Sum of           Mean
      Source                DF     Squares         Square    F Value    Pr > F
      Model                  2       28930          14465       7.68    0.0010
      Error                 64      120503     1882.85220
      Corrected Total       66      149432

               Root MSE             43.39184    R-Square     0.1936
               Dependent Mean       39.10448    Adj R-Sq     0.1684
               Coeff Var           110.96388

                            Parameter Estimates
                           Parameter       Standard
         Variable     DF    Estimate          Error    t Value    Pr > |t|
         Intercept     1   -17.55503       15.95865      -1.10      0.2754
         maxdepth      1     0.48106        0.18111       2.66      0.0100
         no3           1     8.28473        2.95659       2.80      0.0067
```

R code and output

```r
d = read.csv("longnosedace.csv")
m = lm(count~no3+maxdepth,d)
summary(m)
```

```
Call:
lm(formula = count ~ no3 + maxdepth, data = d)

Residuals:
   Min     1Q Median     3Q    Max
-55.06 -27.70  -8.68  11.79 165.31

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -17.555     15.959   -1.10   0.2754
no3            8.285      2.957    2.80   0.0067 **
maxdepth       0.481      0.181    2.66   0.0100 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 43.4 on 64 degrees of freedom
Multiple R-squared:  0.194, Adjusted R-squared:  0.168
F-statistic: 7.68 on 2 and 64 DF,  p-value: 0.00102
```

Interpretation

- Intercept (β_0): The expected count of Longnose Dace when maximum depth and nitrate concentration are both zero is -18.
- Coefficient for maxdepth (β_1): Holding nitrate concentration constant, each cm increase in maximum depth is associated with an additional 0.48 Longnose Dace counted on average.
- Coefficient for no3 (β_2): Holding maximum depth constant, each mg/liter increase in nitrate concentration is associated with an addition 8.3 Longnose Dace counted on average.
- Coefficient of determination (R^2): The model explains 19% of the variability in the count of Longnose Dace.