Multiple regression

STAT 401 - Statistical Methods for Research Workers

Jarad Niemi

Iowa State University

October 19, 2013

Multiple regression

Recall the simple linear regression model is

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

The multiple regression model is

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

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

Interpretation

Model:

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

The interpretation is

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

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).*

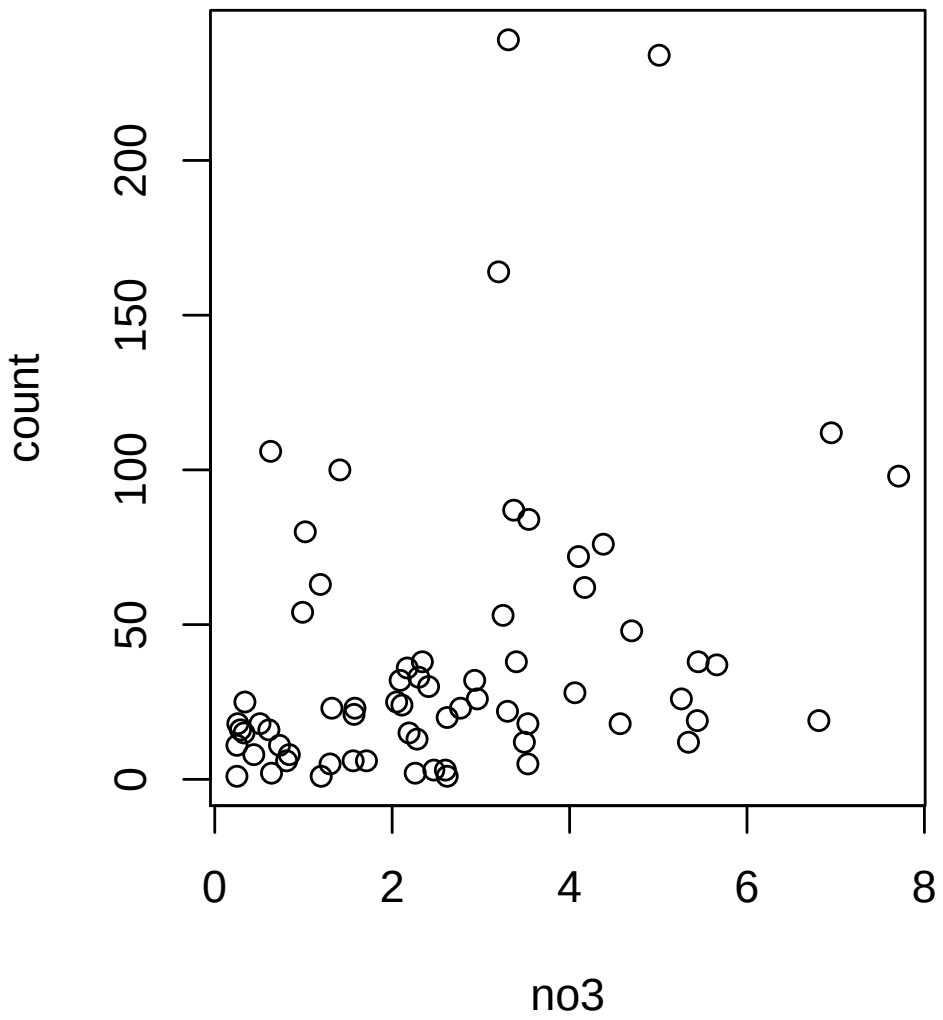
Let's focus on the following model

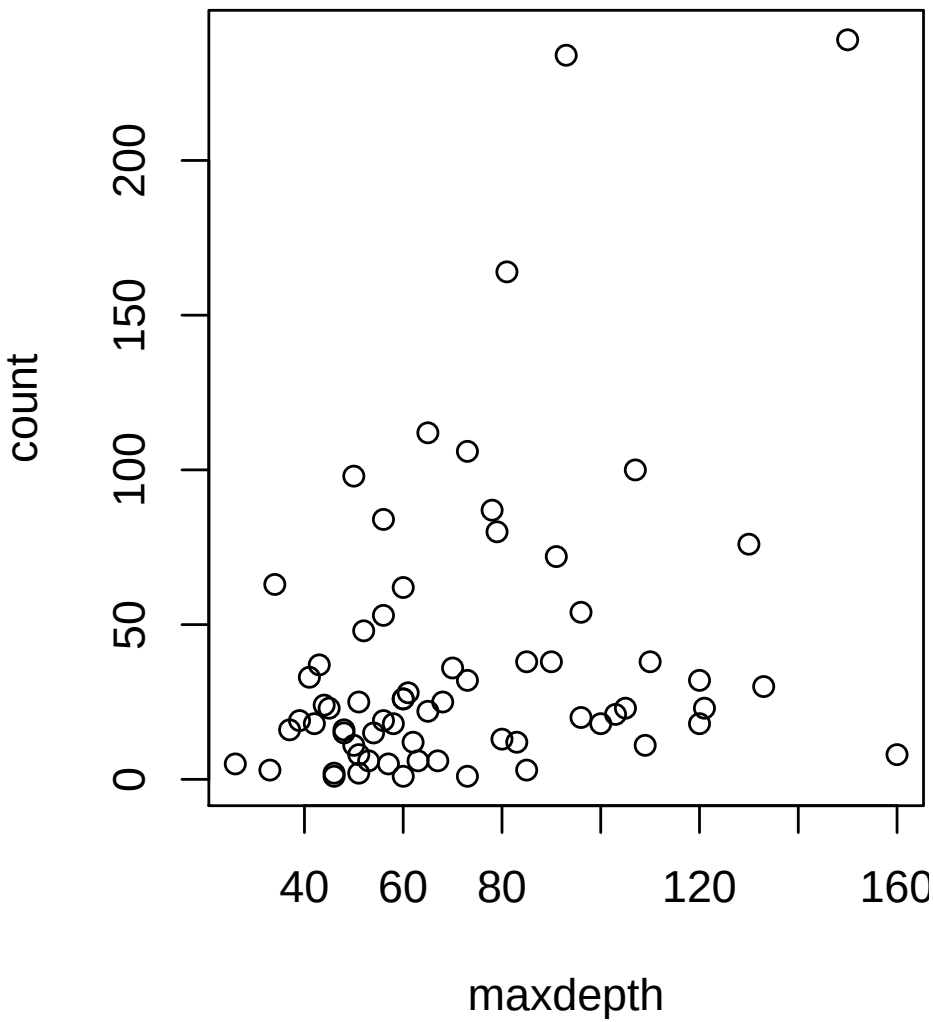
$$Y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2}, \sigma^2)$$

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

Source	DF	Sum of Squares	Mean 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

Variable	DF	Parameter Estimate	Standard 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

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: The model explains 19% of the variability in the count of Longnose Dace.

Future

Possible explanatory variables:

- Additional explanatory variables
- Higher order terms
- Dummy/indicator variables for categorical variables
- Interactions