Multiple regression - indicator functions

STAT 401 - Statistical Methods for Research Workers

Jarad Niemi

Iowa State University

October 20, 2013

Multiple regression

The multiple regression model is

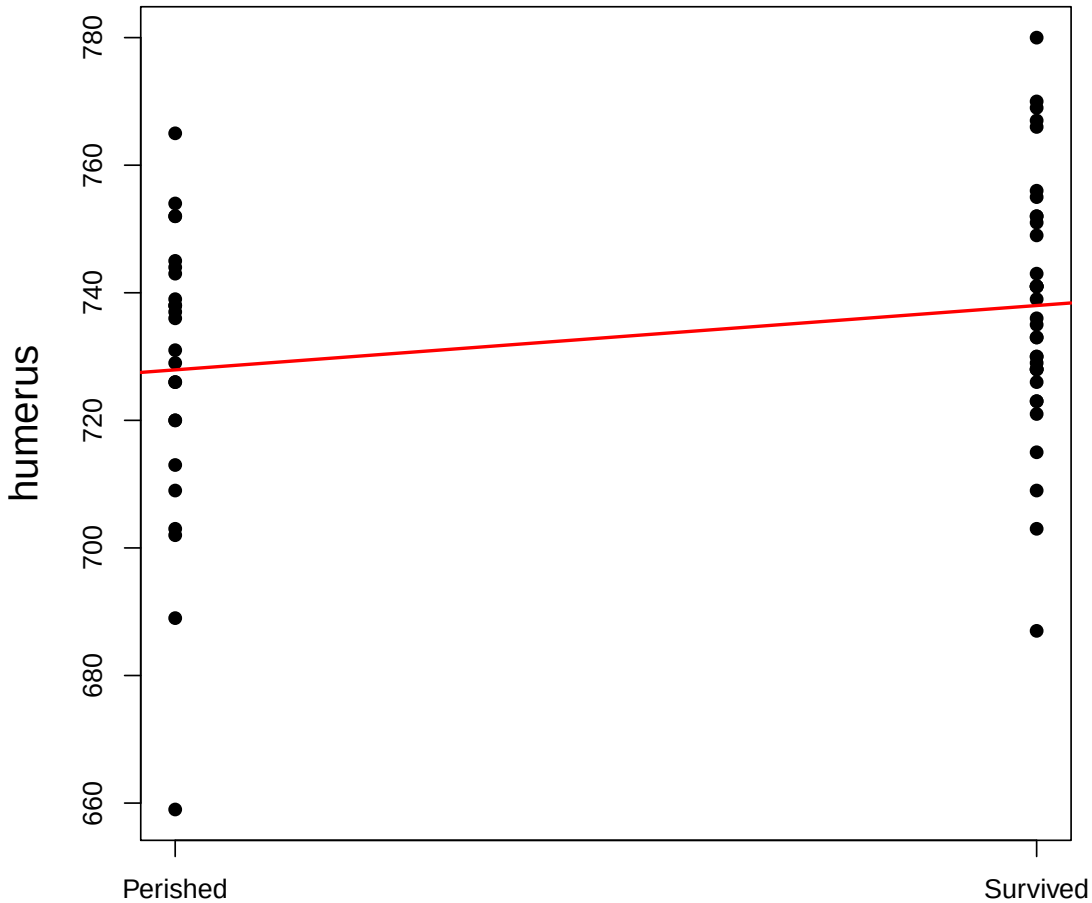
$$Y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p}, \sigma^2)$$

where

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

If we want to incorporate categorical explanatory variables, we need to use indicator functions to construct the explanatory variables.

Two-sample regression



Two-sample regression

- Choose one of the levels as the reference level, e.g. perished
- Construct a dummy variable using an indicator function for the other level, e.g.

$$X_{i,1} = \begin{cases} 1 & \text{observation } i \text{ survived} \\ 0 & \text{otherwise} \end{cases}$$

we often write $X_{i,1} = \mathrm{I}(\text{observation } i \text{ survived})$ where an indicator function has the following definition:

$$\mathrm{I}(\mathrm{A}) = \begin{cases} 1 & \mathrm{A} \text{ is true} \\ 0 & \text{otherwise} \end{cases}$$

- Run a simple linear regression using this dummy variable.

See Section 2.1.1 14:56 Tuesday, February 28, 2012 11

The REG Procedure
Model: MODEL1
Dependent Variable: humerus

Number of Observations Read	59
Number of Observations Used	59

Analysis of Variance

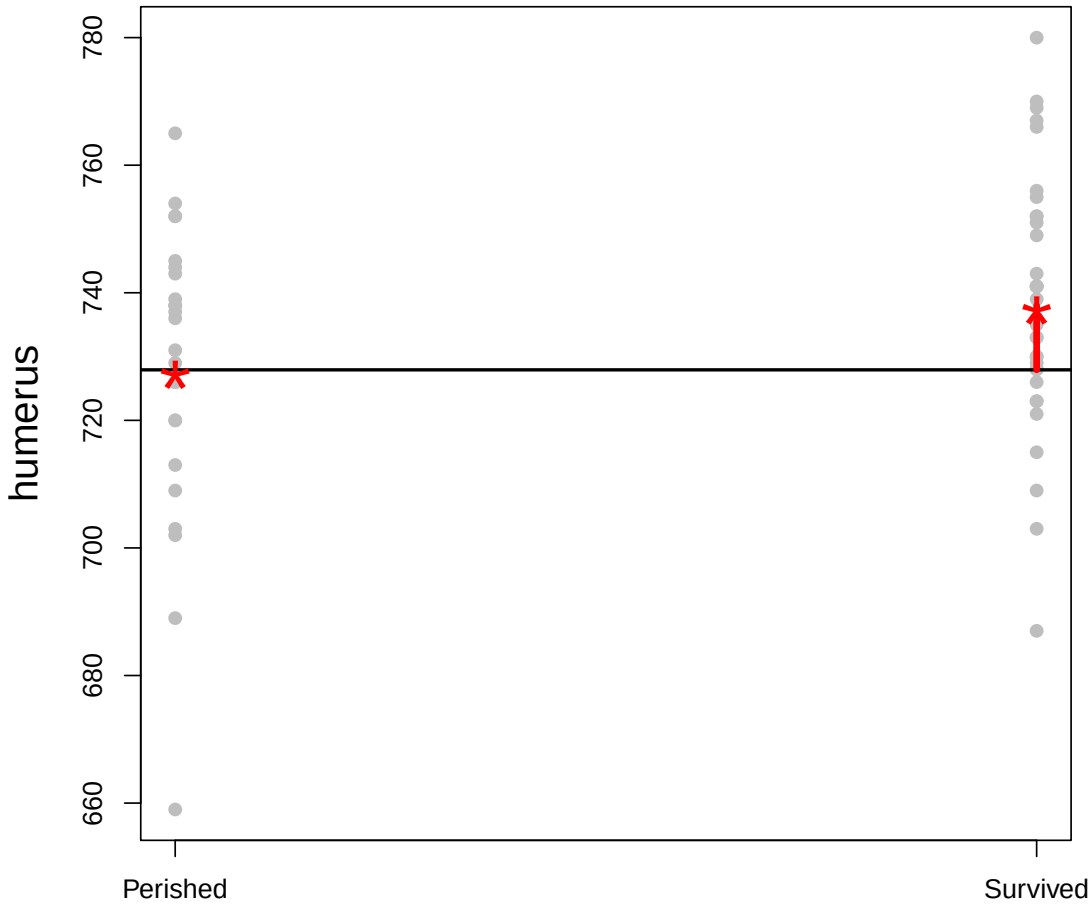
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1447.55650	1447.55650	3.16	0.0809
Error	57	26130	458.41813		
Corrected Total	58	27577			

Root MSE	21.41070	R-Square	0.0525
Dependent Mean	733.89831	Adj R-Sq	0.0359
Coeff Var	2.91739		

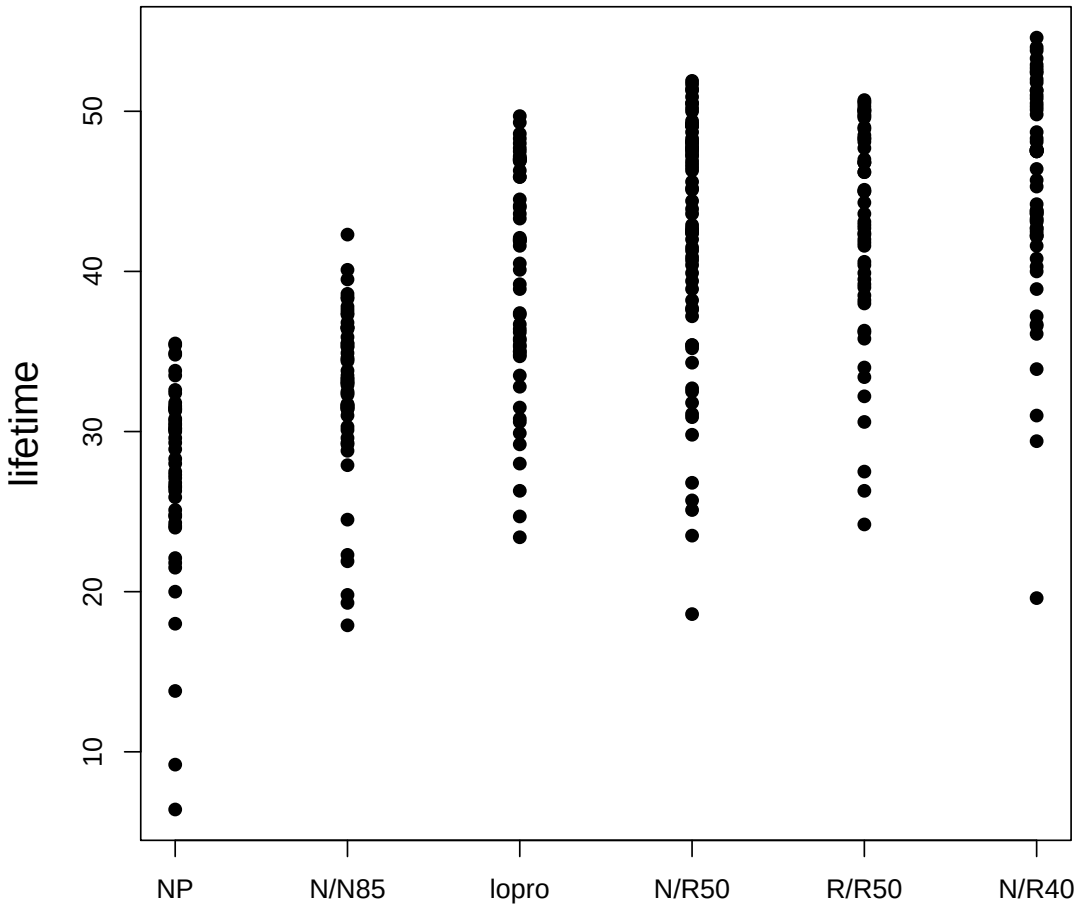
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	727.91667	4.37044	166.55	<.0001
x1	1	10.08333	5.67436	1.78	0.0809

Two-sample regression



Using a categorical variable as an explanatory variable.



Regression with a categorical variable

- Choose one of the levels as the reference level, e.g. N/N85
- Construct dummy variables using indicator functions for the other levels, e.g.

$$
\begin{aligned}
X_{i,1} &= \mathrm{I}(\text{diet for observation } i \text{ is NP}) \\
X_{i,2} &= \mathrm{I}(\text{diet for observation } i \text{ is N/R50 lopro}) \\
X_{i,3} &= \mathrm{I}(\text{diet for observation } i \text{ is N/R50}) \\
X_{i,4} &= \mathrm{I}(\text{diet for observation } i \text{ is R/R50}) \\
X_{i,5} &= \mathrm{I}(\text{diet for observation } i \text{ is N/R40})
\end{aligned}
$$

- Run a multiple linear regression using these dummy variables.

```
DATA case0501;
  INFILE 'U:/401A/Sleuth Datasets/CSV/case0501.csv' DSD FIRSTOBS=2;
  INPUT lifetime diet $;
  IF diet ='NP'    THEN x1=1; ELSE x1=0;
  IF diet ='lopro' THEN x2=1; ELSE x2=0;
  IF diet ='N/R50' THEN x3=1; ELSE x3=0;
  IF diet ='R/R50' THEN x4=1; ELSE x4=0;
  IF diet ='N/R40' THEN x5=1; ELSE x5=0;
  RUN;

PROC REG DATA=case0501;
  MODEL lifetime = x1 x2 x3 x4 x5;
  RUN; QUIT;
```

The REG Procedure
Model: MODEL1
Dependent Variable: lifetime

Number of Observations Read	349
Number of Observations Used	349

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	12734	2546.78836	57.10	<.0001
Error	343	15297	44.59888		
Corrected Total	348	28031			

Root MSE	6.67824	R-Square	0.4543
Dependent Mean	38.79713	Adj R-Sq	0.4463
Coeff Var	17.21323		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	32.69123	0.88455	36.96	<.0001
x1	1	-5.28919	1.30101	-4.07	<.0001
x2	1	6.99449	1.25652	5.57	<.0001
x3	1	9.60596	1.18768	8.09	<.0001
x4	1	10.19449	1.25652	8.11	<.0001
x5	1	12.42544	1.23521	10.06	<.0001

```
DATA case0501;
  INFILE 'U:/401A/Sleuth Datasets/CSV/case0501.csv' DSD FIRSTOBS=2;
  INPUT lifetime diet $;
  IF diet = 'N/N85' THEN diet = 'zN/N85';

PROC GLM DATA=case0501;
  CLASS diet;
  MODEL lifetime=diet / SOLUTION;
  RUN;
```

The GLM Procedure

Dependent Variable: lifetime

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	12733.94181	2546.78836	57.10	<.0001
Error	343	15297.41532	44.59888		
Corrected Total	348	28031.35713			

R-Square	Coeff Var	Root MSE	lifetime Mean
0.454275	17.21323	6.678239	38.79713

Source	DF	Type I SS	Mean Square	F Value	Pr > F
diet	5	12733.94181	2546.78836	57.10	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
diet	5	12733.94181	2546.78836	57.10	<.0001

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		32.69122807	B	0.88455439	36.96	<.0001
diet	N/R40	12.42543860	B	1.23521298	10.06	<.0001
diet	N/R50	9.60595503	B	1.18768248	8.09	<.0001
diet	NP	-5.28918725	B	1.30100640	-4.07	<.0001
diet	R/R50	10.19448622	B	1.25652099	8.11	<.0001
diet	lopro	6.99448622	B	1.25652099	5.57	<.0001
diet	zN/N85	0.00000000	B	.	.	.

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

Multi-sample regression