Data: The Heart of Statistics

Bret Hanlon and Bret Larget

Department of Statistics
University of Wisconsin—Madison

September 2, 2010

Cow Example

Example

A study assigned 50 cows to various diets (based on the amount of an additive in the diet) and examined a number of outcomes associated with characteristics of the produced milk, amount of dry matter consumed, and weight gain of the cow. Pre-treatment variables include initial weight of the cow, number of lactations, and age of the cow. The primary purpose of the study was to examine the effect of the different diets on the outcome variables, controlling for effects of other covariates.

Cow variables

The variables in the data set are:

- **treatment** the diet, one of CONTROL, LOW, MEDIUM, and HIGH;
- **level** mg of additive per kg of feed;
- **lactation** the number of lactations (pregnancies);
- **age** age of the cow at the beginning of the study in months;
- **initial.weight** initial weight in pounds;
- **dry** mean daily weight of dry matter consumed (kg);
- **milk** mean daily amount of milk produced (pounds);
- **fat** percentage milk fat (grams of fat per 100g milk)
- **solids** percentage of solids in milk by weight (grams solids per 100g milk)
- **final.weight** final weight of cow in pounds;
- **protein** percentage of protein in milk by weight (grams protein per 100g milk)

Data

Here is a representative sample of the data:

treatment	level	lactation	age	initial.weight	dry	milk	fat	solids	final.weight	protein
control	0	3	49	1360	15.429	45.552	3.88	8.96	1442	3.67
control	0	3	47	1498	18.799	66.221	3.40	8.44	1565	3.03
control	0	2	36	1265	17.948	63.032	3.44	8.70	1315	3.40
control	0	2	33	1190	18.267	68.421	3.42	8.30	1285	3.37
control	0	2	31	1145	17.253	59.671	3.01	9.04	1182	3.61
control	0	1	22	1035	13.046	44.045	2.97	8.60	1043	3.03
low	0.1	6	89	1369	14.754	57.053	4.60	8.60	1268	3.62
low	0.1	4	74	1656	17.359	69.699	2.91	8.94	1593	3.12
low	0.1	3	45	1466	16.422	71.337	3.55	8.93	1390	3.30
low	0.1	2	34	1316	17.149	68.276	3.08	8.84	1315	3.40
low	0.1	2	36	1164	16.217	74.573	3.45	8.66	1168	3.31
low	0.1	2	41	1272	17.986	66.672	3.43	9.19	1188	3.59
medium	0.2	3	45	1362	19.998	76.604	4.29	8.44	1273	3.41
medium	0.2	3	49	1305	19.713	64.536	3.94	8.82	1305	3.21
medium	0.2	3	48	1268	16.813	71.771	2.89	8.41	1248	3.06
medium	0.2	3	44	1315	15.127	59.323	3.13	8.72	1270	3.26
medium	0.2	2	40	1180	19.549	62.484	3.36	8.51	1285	3.21
medium	0.2	2	35	1190	19.142	70.178	3.92	8.94	1168	3.28
high	0.3	5	81	1458	20.458	71.558	3.69	8.48	1432	3.17
high	0.3	3	49	1515	19.861	56.226	4.96	9.17	1413	3.72
high	0.3	3	48	1310	18.379	49.543	3.78	8.41	1390	3.67
high	0.3	3	46	1215	18.000	55.351	4.22	8.94	1212	3.80
high	0.3	3	49	1346	19.636	64.509	4.16	8.74	1318	3.31
high	0.3	3	46	1428	19.586	74.430	3.92	8.75	1333	3.37

Categorization of variables

- Variables are (usually) either *numerical* (quantitative) or *categorical* (qualitative).
  - numerical variables take on numerical values and are either *discrete* or *continuous*.
  - categorical variables partition the observations into categories: if the categories have a natural order, the variable is *ordinal*; if not, it is *nominal*.
- Variables are *experimental* or *observational*.
  - experimental variables have values that are under control of the researcher.
  - observational variables have values that are observed and are not set by the researcher.
- Variables may be *response variables* or *explanatory variables*.
  - response variables are considered as outcomes;
  - explanatory variables are thought potentially to affect outcomes.

Classify each of the variables in the cow example.

Level of Measurement

- Data is often represented in a rectangular array where each column is a *variable* and each row is an *observational unit* or *sampling unit*.
- In the cow example, individual cows are sampling units and each variable measures something on the level of the cow (although in some sense, the quantitative variable LEVEL measures something on the level of the treatment).
- In other examples, there may be multiple levels of measurement.
- It is important to recognize different levels of measurement because it can affect the selection of an approriate method of analysis.

Plantation Data

Example

Researchers interested in forest restoration in Costa Rica conducted an experiment to examine which of several species of tree best promoted the growth of native woody plants in their understory in plantations that were being converted back to natural forest. The approach was to plant a fast-growing native tree in the plantation that would provide shade and a suitable environment for additional native species to become established. At some point, the planted overstory trees would be harvested, leaving a diverse natural forest behind.

Plantation Data (continued)

Example

The study included three sites, each a plantation that had been previously cleared for agriculture. One site (La Selva) was a former experimental research station, while the other two sites (Paniagua and Quesada) had been private farms. The La Selva, Paniagua, and Quesada plantations were 100 m, 1.3 km, and 2.5 km from continuous forest, respectively. Each site was divided into six plots (of various sizes) and each plot was planted with one of six species of tree, spaced in a regular array (of varying sizes). With minimal management, the sites were allowed to grow for nearly a decade. Each plot included four subplots (4m by 4m) for which several variables were measured. The primary response variable is the number of woody stemmed plants in each subplot. Other variables include the percentage of the subplot shaded by the canopy of the overstory, whether or not the subplot was flat or sloped, and whether or not the subplot had good drainage.

Sample Data

stems	canopy	site	overstory	spacing	slope	drainage	distance
250	14	LaSelva	Cb	8	flat	good	0.1
60	15	LaSelva	Cb	8	flat	good	0.1
46	13	LaSelva	Ta	8	flat	good	0.1
36	15	LaSelva	Ta	8	flat	good	0.1
125	14	LaSelva	Vg	8	flat	good	0.1
110	13	LaSelva	Vg	8	flat	good	0.1
45	12	LaSelva	Ha	8	flat	good	0.1
50	13	LaSelva	Ha	8	flat	good	0.1
10	15	Paiagua	Cb	4	sloped	good	1.3
0	11	Paiagua	Cb	4	sloped	good	1.3
26	14	Paiagua	Ta	32	flat	good	1.3
30	15	Paiagua	Ta	32	flat	good	1.3
22	10	Paiagua	Vg	64	sloped	good	1.3
15	10	Paiagua	Vg	64	sloped	good	1.3
4	8	Paiagua	Ha	32	flat	poor	1.3
25	8	Paiagua	Ha	32	flat	poor	1.3
22	11	Quesada	Cb	16	flat	poor	2.5
30	11	Quesada	Cb	16	flat	poor	2.5
33	12	Quesada	Ta	16	flat	good	2.5
30	11	Quesada	Ta	16	flat	good	2.5
40	23	Quesada	Vg	36	flat	good	2.5
4	24	Quesada	Vg	36	flat	good	2.5
0	10	Quesada	Ha	16	flat	poor	2.5
2	10	Quesada	Ha	16	flat	poor	2.5

Variables

- Each row of the data set represents a subplot (many are not shown) and each column a variable.
- For each variable, categorize it (numerical or categorical; response or explanatory; experimental or observational).
- For each variable, identify the sampling unit.

stems is the number of woody stems;
canopy is the percentage of overstory canopy;
site is the name of the plantation;
overstory is an abbrviation of the species of planted tree;
spacing is the number of square meters per planted tree;
slope is either FLAT or SLOPED;
drainage is either GOOD or POOR;
distance is the distance (km) to the nearest continuous forest.

R Demonstration

- Demonstrate reading data into R and do some simple graphs and numerical summaries.

What you should know

You should know:

- how to distinguish between different types of variables;
- how to determine possible values for each variable;
- how to determine the sampling unit for each variable.