Kerala School of Mathematics
Course in Statistics for Scientists
Regression

T.Krishnan
Strand Life Sciences, Bangalore

6 March 2010

Example: Arsenic in Water and in Nails

- Toenail samples are indicators of drinking water arsenic exposure.
- Arsenic accumulates in bones, hair, and nails and is used to detect chronic exposure, since arsenic is laid down in keratin soon after ingestion.
- The analysis of trace elements in biological samples will extend our understanding of the impact that environmental exposure to these elements has on human health.
- Measuring arsenic content in nails has proven useful in studies evaluating the chronic body burden of arsenic.

Data Details

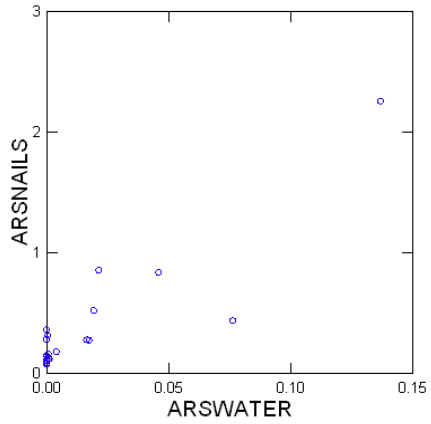
This data set contains measurements of drinking water and toenail levels of arsenic, as well as related factors, for 21 individuals with private wells in New Hampshire, USA.
Variable names in order from left to right:
AGE: Age (years)
GENDER: Gender of person (1 = Male, 2 = Female)
DRINKUSE: Household well used for drinking
$(1 \leq \frac{1}{4}, 2 = \frac{1}{4}, 3 = \frac{1}{2}, 4 = \frac{3}{4}, 5 \geq \frac{3}{4})$
COOKUSE: Household well used for cooking
$(1 \leq \frac{1}{4}, 2 = \frac{1}{4}, 3 = \frac{1}{2}, 4 = \frac{3}{4}, 5 \geq \frac{3}{4})$
ARSWATER: Arsenic in water (ppm)
ARSNAILS: Arsenic in toenails (ppm)

Data

AGE	GENDER	DRINKUSE	COOKUSE	ARSWATER	ARSNAILS
44	f	5	5	0.00087	0.119
45	f	4	5	0.00021	0.118
44	m	5	5	0.00000	0.099
66	f	3	5	0.00115	0.118
37	m	2	5	0.00000	0.277
45	f	5	5	0.00000	0.358
47	m	5	5	0.00013	0.080
38	f	4	5	0.00069	0.158
41	f	3	2	0.00039	0.310
49	f	4	5	0.00000	0.105
72	f	5	5	0.00000	0.073
45	f	1	5	0.04600	0.832
53	m	5	5	0.01940	0.517
86	f	5	5	0.13700	2.252
8	f	5	5	0.02140	0.851
32	f	5	5	0.01750	0.269
44	m	5	5	0.07640	0.433
63	f	5	5	0.00000	0.141
42	m	5	5	0.01650	0.275
62	m	5	5	0.00012	0.135
36	m	5	5	0.00410	0.175

Scatterplot

- Let us plot arsenic in nails against arsenic in water to get an idea of the relationship.
- This plot is called a **Scatterplot**.



- There seems to be an increasing relationship, maybe linear.
- If we had more data we would do this plot and the analysis separately for different factors such as gender, drinkuse, and cookuse.

Prediction of Arsenic Levels in Nails Based on Arsenic in Water

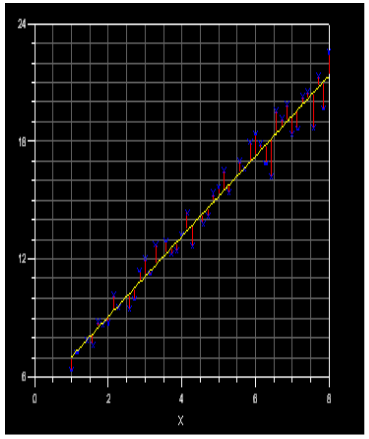
- The data that we have (called **Training Data**) could be used to develop a relationship between Arsnails and Arswater to help predict Arsnails from Arswater at a given level.
- Such an exercise is called **Regression Analysis**.
- The simplest type of such a relationship is **linear**, that is a straight line.
- The data points will in general not lie on a straight line.
- The formula developed finds the straight line that minimizes the prediction error.
- For this purpose it uses a technique called **least squares**.
- Using the data it is also possible to develop a measure of the goodness of this prediction formula.
- If the data points lie very close to a straight line the prediction formula is expected to be good.

What is Regression?

- A body of statistical methods for finding a predictive formula (often a line or surface of line of “best fit”) for one response (dependent) numerical variable based on one or more explanatory (independent) variables or predictor variables.
- These methods include assess of the "goodness of fit" of the formula, (for example, the Correlation Coefficient).
- Data required for developing the formula are on the response as well as on the predictor variables, called “training data”.
- The formula developed minimizes the sum of squares of the distances of the data points from the fitted line or surface, **along the response (y) axis**.
- Along the y-axis because the error is between the y-value observed and that predicted by the straight line.
- The linear regression computed in this way finds the linear function of the predictor variables which has the largest correlation among all linear functions of the predictor variables.

Least Squares

- The figure indicates the distance between the data point and the vertical distance (along y-axis) to the straight line.
- This is a different exercise from curve-fitting where the perpendicular distance of the data point to the straight line is of importance.



Regression Results for Arsenic Data

Regression Coefficients B = $(X'X)^{-1}X'Y$

Effect	Coefficient	Standard Error	Std. Coefficient	Tolerance	t	p-value
CONSTANT	0.155	0.054	0.000	.	2.875	0.010
ARSWATER	12.986	1.473	0.896	1.000	8.816	0.000

Confidence Interval for Regression Coefficients

Effect	Coefficient	95.0% Confidence Interval		VIF
		Lower	Upper	
CONSTANT	0.155	0.042	0.268	.
ARSWATER	12.986	9.903	16.068	1.000

Correlation Matrix of Regression Coefficients

	CONSTANT	ARSWATER
CONSTANT	1.000	
ARSWATER	-0.445	1.000

Analysis of Variance

Source	SS	df	Mean Squares	F-ratio	p-value
Regression	3.809	1	3.809	77.724	0.000
Residual	0.931	19	0.049		

Interpretation of Regression Results

- The regression line is

$$\text{arsnails} = 0.155 + 12.986 \times \text{arswater}.$$

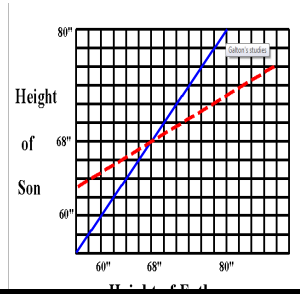
- The correlation coefficient between arsnails and arswater is 0.896, which is reasonably high indicating a fairly strong linear relationship.
- This formula can be used to predict arsnails value at say arswater=0.01 as

$$\text{arsnails prediction} = 0.155 + 12.986 \times 0.01 = 0.275.$$

- It is not a good idea to use this formula beyond the region of your data.
- The output gives standard errors, confidence intervals, significance tests, etc. for the regression coefficients.
- The ANOVA table shows that the regression is significant meaning thereby that the arswater information is useful in predicting arsnails.

Why Is It Called **Regression**?

- Galton noticed that extremely tall fathers tended to have sons shorter than themselves, and extremely short fathers tended to have sons taller than themselves.
- "Tallness" or "shortness" did not breed true like they did in Mendel's pea experiments.
- The offspring seemed to regress to the median, or "mediocrity".
- This is shown in the figure below.
- The red line is what is observed. The blue line is what would have been if son's height is equal to his father's.



Some Comments on Use of Regression

- No cause-effect relationship is to be inferred even if the formula is good.
  1. For instance, weekly growth of grass in a lawn and the average depth of the local reservoir may be highly correlated and one may be well predictable from the other, but one does not cause the other. It is just that both are dependent upon a third variable, the rainfall.
  2. It requires a controlled experiment to establish cause-effect relationship.
- Do not predict outside the data range.
  1. In the arsenic example if you use the formula for arswater= 0.5 then the prediction of arsnails is 6.65, an untenable figure.
  2. The linear regression model is unlikely to hold far away from the data range.
- Regression analysis provides standard error for prediction and it should be computed and quoted along with prediction, or confidence intervals for prediction is to be mentioned in results.
- These ideas are extendable to several predictor (independent) variables—called **Multiple Regression**.
- Regression coefficients can also be used for explaining phenomena, not only for prediction.
- Polynomials and transformed variables (like logarithm) may also be used in the linear regression set-up.
- Nonlinearity in predictors need more complicated techniques.
- Regression analysis should be accompanied by regression diagnostics to study the effect of outliers and special observations on the regression line.