Linear Regression

LECTURE 17

Objectives

- Define terms.
- Describe linear regression.
- Describe the use and limitations of linear regression
- Explain the assumptions of linear regression.
- Perform linear regression.

Overview

- Up to this point all of our independent variables have been nominal data (male vs. female, site A vs. site B, high concentration vs. low concentration).
- Independent variable does not have to be nominal data – they can also be interval or ratio scale data.

Overview

- In univariate statistics only a couple of techniques are appropriate when you have an independent variable which is ratio or interval.
  - Correlation
  - Linear regression

Overview

- Correlation and regression can use a ratio/interval independent variable without recoding the data:
  - They allow for greater power.
  - They allow for greater predictive accuracy.

Overview

- Linear regression
  - Establishes a causal relationship between the independent and dependent.
  - Allows for the value of the dependent variable to be predicted if the value of independent variable is known.
  - System needs to be well understood to suggest causation – hypothesis testing.
  - Statistical assumptions are less strenuous than those for correlation.

Overview

- Intent
  - Establish a causal relationship between the independent and dependent variable.
  - Allow for the prediction of the value of dependent variable given the value for the corresponding independent variable.
  - Linking the results of statistical analyses to graphical interpretation.

Overview

- Data requirements
  - Independent variable – One interval/ratio
    - Fixed variable.
    - Random variable.
  - Dependent variable – One interval/ratio

Overview

- Data requirements
  - Minimum data requirements: None specified.
    - Linear regression can be conducted on as few as two data points.
    - Predictive ability and power improves with increasing sample size.
    - Lower reasonable limit is 6 to 8 samples.

Basic linear equations

$$Y_i = \alpha + \beta X_i$$

- Yi represents the value of the line on the y-axis at the corresponding Xi
- α is the y-axis coordinate for the y-intercept.
- β is the slope of the line

Basic linear equations

- Because the points rarely fall exactly on the line another term is frequently added.

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

  - ε_i represents the error or portion of Y_i that can not be explained by X_i. Frequently referred to as a residual.

Rationale

- Suppose we suspect that the size of a female snake influences the number of eggs that she lays.
  - We collect data on the size of female snakes and the number of eggs that the female laid.

Female size (inches)	Number of eggs laid
12	2
13	2
14	3
15	4
16	3
17	4
18	6

Rationale

- This still says very little on the relationship between size and number of eggs laid.
  - In fact, at this point we can do little more than guess that regardless of body size each snake will produce the mean number of eggs.

Female size	Number of eggs	Mean number of eggs
12	2	3.4286
13	2	3.4286
14	3	3.4286
15	4	3.4286
16	3	3.4286
17	4	3.4286
18	6	3.4286

Rationale

- This guess is obviously not correct and we can calculate how far our guess was from the truth.

Female size	Number of eggs	Mean eggs	Difference
12	2	3.4286	-1.4286
13	2	3.4286	-1.4286
14	3	3.4286	-0.4286
15	4	3.4286	0.5714
16	3	3.4286	-0.4286
17	4	3.4286	0.5714
18	6	3.4286	2.5714
Total			**0**

Rationale

- This conveys more information because it shows the deviation of each datum from the mean, but is does not indicate how our guess did overall of estimating the true number of eggs produced, because the positive values are cancelled by the negative values. To counter this we can square the differences.

Size	Eggs	Mean eggs	Difference	Difference squared
12	2	3.4286	-1.4286	2.0409
13	2	3.4286	-1.4286	2.0409
14	3	3.4286	-0.4286	0.1837
15	4	3.4286	0.5714	0.3265
16	3	3.4286	-0.4286	0.1837
17	4	3.4286	0.5714	0.3265
18	6	3.4286	2.5714	6.6121
Total			**0**	11.7143

Rationale

- Total of the differences squared is referred to as the Sum of the Squares.
- An effective measure of how good our guess (the mean) estimates the actual values.
  - The smaller the sum of squares the better the estimate approximates the actual values.
  - The large the sum of squares the worse the estimate approximates the actual values.

Rationale

- At this point we need to clarify what we mean by relationship:
  - Do we mean "How do eggs and body size covary?"
  - Do we mean "Can we use size to predict the number of eggs that were laid?"

Rationale

- The key to regression analysis is to find a linear equation that minimizes the sum of the squares (allows us to make as accurate predictions as possible).

Rationale

- The slope is frequently referred to as the regression coefficient.
- The Y-intercept is referred to as the intercept or constant.
- The variation that cannot be explained by the regression line is referred to as residual or error variation. (Y_i-Y^)

Rationale

- The amount of the actual variation in the data set explained by the linear regression is expressed as a proportion (0-1) and is called the coefficient of determination (r^2).

Rationale

- The significance of the regression
  - Compare the predicted Y values from the regression equation with the actual Y values and the values of Y predicted from a linear equation with a slope of 0 passing through the Y-mean.
- This comparison is typically done as a variation of an ANOVA.

Assumptions

- Random sampling
- Independent observations
- The value of the independent variable determines (causes) the value of the dependent variable.
- The relationship between the two variables is linear.
- The X-variable is under the control of the observer and the X-values are assumed to be exact.
- For any value of X there is a theoretical population of Y-values that are normally distributed.
- The variances of Y are equal (Homoscedasticity).

Usage rules

- Most relationships in biology are not linear.
- Interpolating and extrapolating

Example

- Example 21