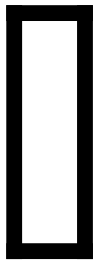

Inferential Statistics: Hypothesis Testing and Regression Modeling

CPS 5310 Spring 2014

Shirley Moore, Instructor

February 4, 2014

Learning Objectives

- After completing this lesson (including homework), you should be able to
  - Test hypotheses for one-sample and two-sample datasets, assuming normal distribution of the data using R's t.test function
  - Perform single-classification analysis of variance using R's anova function
  - Perform simple and multiple linear regression using R's lm function
  - Perform nonlinear regression using R's nls function
  - Evaluate the goodness of fit of a regression model using the coefficient of determination
  - Evaluate the goodness of prediction of a regression model using cross-validation

Inferential Statistics

- Used to draw *inferences* from data – i.e., make generalizations or extrapolate or draw conclusions from statistical sample data
- Methods
  - hypothesis testing
  - regression
    - linear
    - nonlinear
    - multiple

Example 1: Crop Yields

- Crop A: 715, 683, 664, 659, 660, 762, 720, 715; mean 697.25
- Crop B: 684, 655, 657, 531, 638, 601, 611, 651; mean 628.5
- Data in PhenMod/Stat/crop.csv
- Is Crop A's yield really higher or is the difference just a random effect?
- Questions of this nature can be answered by statistical hypothesis testing.

Hypothesis Test Steps

1. Select the hypothesis to be tested, the *null hypothesis* H_0.
2. May need to select alternative *hypothesis* H_1 that is assumed to hold true if H_0 is rejected. Let H_1 be the negation of H_0 if no alternative hypothesis is selected.
3. Select the significance level α, which is the probability of erroneously rejecting a true H_0 as a result of the test. Typical values for α are between 0.01 and 0.1.
4. Collect the necessary data and use software to perform the test which will give you a p value.
5. If $p < \alpha$, reject H_0. In this case, H_1 is assumed to hold true, and H_1 is said to be *statistically significant at level* α. If $p \geq \alpha$, nothing can be derived from the test.

Example 1: Crop Yields (cont.)

- Let X_1 and X_2 be the random variables that generated the data for crop A and crop B, respectively, and let μ_1 and μ_2 denote the (unknown) expected values of these random variables. Define the data for the statistical test as follows:
  - H_0: $\mu_1 = \mu_2$
  - H_1: $\mu_1 > \mu_2$
  - $\alpha = 0.05$
- Use the t.test in R to get a *p* value
  - t.test(dataset$x, dataset$y, alternative="greater", paired=FALSE)
  - t.test assumes normal distributions of the random variables and of the differences between them.

Example 2. Two-sample t-test: Car Mileage

- http://www.r-tutor.com/elementary-statistics/inference-about-two-populations/population-mean-between-two-independent-samples

Example 3. One-Sample t-test: Normal Body Temperature

- What is normal human body temperature (taken orally)? We've all been taught since grade school that it's 98.6 degrees Fahrenheit, and never mind that what's normal for one person may not be "normal" for another! So from a statistical point of view, we should abandon the word "normal" and confine ourselves to talking about mean human body temperature. We hypothesize that mean human body temperature is 98.6 degrees, because that's what we have been tols. The data set "Normal Body Temperature, Gender, and Heart Rate" bears on this hypothesis. The data are from a random sample (supposedly) of 130 cases and are posted at the Journal of Statistical Education's data archive. The original source of the data is Mackowiak, Wasserman, and Levine (1992). A Critical Appraisal of 98.6 Degrees F, the Upper Limit of Normal Body Temperature. Journal of the American Medical Association, 268, 1578-1580.
- Data file: normtemp.txt

t.test in R

t.test(x, y = NULL, alternative = c("two.sided", "less", "greater"), mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95)

- x is a numeric vector of data values and y is an optional numeric vector of data values. If y is excluded, the function performs a one-sample t-test on the data contained in x, if it is included it performs a two-sample t-tests using both x and y.
- The option *mu* provides a number indicating the true value of the mean (or difference in means if you are performing a two sample test) under the null hypothesis.
- The option *alternative* specifies the alternative hypothesis, and must be one of the following: "two.sided" (which is the default), "greater" or "less" depending on whether the alternative hypothesis is that the mean is different than, greater than or less than mu, respectively.
- The option *paired* indicates whether or not you want a paired t-test. The default is FALSE and means that x and y are independent.
- The option *var.equal* is a logical variable indicating whether or not to assume the two variances are equal when performing a two-sample t-test.
- The option *conf.level* determines the confidence level of the reported confidence interval for μ in the one-sample case and $\mu_1 - \mu_2$ in the two-sample case.

Example 4: Paired t-test

- Wiebe and Bortolotti (2002) examined color in the tail feathers of northern flickers. Some of the birds had one "odd" feather that was different in color or length from the rest of the tail feathers, presumably because it was regrown after being lost. They measured the yellowness of one odd feather on each of 16 birds and compared it with the yellowness of one typical feather from the same bird. What was their conclusion? (data in file yellowness.txt)

Reference: Wiebe, K.L., and G.R. Bortolotti. 2002. Variation in carotenoid-based color in northern flickers in a hybrid zone. Wilson Bull. 114: 393-400.

Analysis of Variance (ANOVA)

- Technique for comparing the means of groups of measurement data
- One-way anova (also called single-classification anova)
  - Multiple observations of the measurement variable are made for each value of the *nominal variable* (also called the *factor*)
  - Example: Investigate whether different fungicides (A, B, C, No Fungicide) have a significantly different impact on the density of fungal spores (Data file: PhenMod/Stat/fungicide.csv)
- Null hypothesis is that the means of the measurement variables are the same for the different categories of data.
- Alternative hypothesis is that the means are not the same.
- Book software: PhenMod/Stat/Anova.R

Regression Models

- The dependent variable is expressed in terms of the independent variables using various types of regression expressions.
- Parameters in the regression equations are tuned to fit the equations to the data.
- Types of regression:
  - Linear regression (dependent variable depends linearly on the regression coefficients)
    - with linear functions
    - with nonlinear functions
  - Multiple linear regression (multiple independent variables)
  - Nonlinear regression (dependent variable depends nonlinearly on the regression coefficients)
  - Multiple nonlinear regression

Linear Regression

- $y = \alpha + \beta x + \varepsilon$ modeled by $\hat{y} = ax + b$
- a and b are the regression coefficients (also called regression parameters)
- x is the explanatory variable (or independent variable)
- y is the response variable (or dependent variable)
- Goal: Determine coefficients that minimize the residual sum of squares RSQ.

$$RSQ = \sum_{i=1}^{m} (y_i - \hat{y}(x_i))^2$$

- Resulting regression equation can be used to predict values of the response variable for values of the explanatory variable near the data.

Linear Regression Examples

- Book software PhenMod\LinReg\LinRegEx1.r
- www.r-tutor.com/elementary-statistics/simple-linear-regression

Coefficient of Determination

- To judge the quality of fit between the model and the data
  - Look at graphical comparison of model to data
  - Compute *coefficient of determination (R^2)*
- *Coefficient of determination*
  - Measures quality of fit between model and data on a scale of 0 to 100%
  - Ratio of the variance of the predicted values to the variance of the predicted values
  - R^2 = P% is interpreted as "P percent of the variance in the measurement data is explained by the model."

(Nonlinear) Linear Regression

- General form $\hat{y}(x) = a_0 + a_1 f_1(x) + a_2 f_2(x) + \cdots + a_s f_s(x)$
- Example
  - Concentration of GAG in urine of children aged 0 to 17
  - LinRegEx4.r (simple linear regression)
  - LinRegEx5.r (polynomial regression function0

Multiple Linear Regression

- As a linear function of multiple explanatory variables

$$\hat{y}(x_1, x_2, \ldots, x_n) = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_n x_n$$

- General form

$$\hat{y}(\vec{x}) = a_0 + a_1 f_1(\vec{x}) + a_2 f_2(\vec{x}) + \cdots + a_s f_s(\vec{x})$$

where $\vec{x} = (x_1, x_2, \ldots, x_n)^t$ and the f_i are arbitrary real functions.

Multiple Linear Regression Example

- Degree of wilting of roses depending on concentrations of certain carbohydrates
- Book software
  - Volz.csv
  - LinRegEx3.r

Cross-Validation

- Partition data set into
  - training dataset used to obtain the regression equation
  - test dataset used to assess the regression equation's predictive capability
- Example using rose wilting data
  - LinRegEx3.r

Nonlinear Regression

- Regression equations depend in a nonlinear way on the regression coefficients
- Use R's nls function
- Example: economic cycle
  - klein.csv
  - NonRegEx1.r

Multiple Nonlinear Regression

- Example: stormer viscometer
  - Measures the viscosity of a fluid by measuring the time taken for an inner cylinder to perform a fixed number of revolutions in response to an actuating weight
  - Calibrating by measuring the time taken with varying weights using fluids of known viscosity

$$T = \frac{a_1 v}{w - a_2}$$

  - Once *a1* and *a2* are determined, the equation can be used to determine the viscosity *v* from measured values of *w* and *T*.
  - Book software
    - stomer.csv
    - NonRegEx2.r