Biostatistics

Lecture10

Regression and Correlation Methods

Simple Linear Regression

- ANOVA was extension of T-Test to multiple group means
- Linear regression extends ANOVA to continuous predictor variables
  - Systolic blood pressure predicted by body mass index
  - Body mass index predicted by caloric intake
  - Caloric intake predicted by measure of stress
    - However, may relate to inactivity due to lack of visual acuity

Simple Linear Regression

- Important to specify a biologically plausible model
  - Systolic blood pressure predicted by eye color
  - Body mass index predicted by visual acuity
    - However, may relate to inactivity due to lack of visual acuity

Regression Models - I

- Assumptions are important in linear regression, but are not absolute
  - Predictor variables are ‘fixed’; i.e., same meaning among individuals
  - Predictor variable measured ‘without error’
  - For each value of the predictor variable, there is a normal distribution of outcomes (subpopulations) and the variance of these distributions are equal

Regression Models - I

- Assumptions (continued)
  - The means of the outcome subpopulations lie on a straight line related to the predictors; i.e., the predictors and the outcomes are linearly related
    - $\mu_{y|x} = \alpha + \beta x$
  - The outcomes are independent of each other
- Regression model: $y = \alpha + \beta x + \varepsilon$

Interpretation of regression model

- Two parameters
  - α is the intercept (Y value) when the predictor is zero
    - May not be really plausible
    - Frequently ‘center’ the data so that ✂ is the value when the predictor is at its mean
  - β is the ‘slope’ of the regression line and represents the change in Y for a unit change in X
    - i.e., a slope of 0.58 would indicate that for a one unit change in X, there is a 0.58 unit change in Y
  - ε is the error term for each individual and is the residual for that individual
    - Residual is the difference between the fitted line (predicted value) and the observed value

Least squares fit

- Regression parameters (α + β) are determined using method of least squares
  - Minimizes the squared differences between each observation and the ‘fitted’ line
  - Can be dramatically affected by unusual values – ‘outliers’

Approach to developing a regression model

- Determine outcome and plausible predictors
- Plot outcome vs. each predictor to check for linearity
- Fit the regression model and review parameters and tests
- If model has a significant fit and parameters are significantly different from 0, look at residuals to better evaulate fit

Anatomy of Regression Table

- Global fit of the model
  - Determine if there are any parameters that are significantly different from zero and, thus, explain some part of the variation
- Coefficient of determination (r^2)
  - Ratio of regression sums of squares (SSR) to the total sums of squares
  - Can be interpreted as the proportion of the total variation in the outcome that can be explained by the regression model

Anatomy of Regression Table

- Total Sums of Squares (SST)
  - SST is total SS in the outcome for the entire sample – analogous to the variance
- Regression Sums of Squares (SSE)
  - Represents the variability in the outcome accounted for by differences among the group means
  - This is the part of the Regression Table that is of most interest
    - It tells you if the factor that you are interested in explains a significant amount of the variance in the sample
    - If the $\mu_{y|x}$ change little across the range of the x's, they will explain very little of the variance in the sample and that predictor (x) will not have a significant F-test in the table (non-significant p-value)

Correlations

- Bivariate correlation (ρ) is an indicator of the strength and direction of the relationship between two variables
  - Related to the slope for the two variables through variance terms
- Correlation ‘matrix’ is frequently used to screen for important relationships
- Ranges from –1 (perfect inverse relationship) to +1 (perfect positive relationship)

Types of Correlations

- Parametric correlations – Pearson (requires normal distribution)
- Nonparametric alternatives – Spearman’s or Kendall’s
- Test of significance – t-test
- Test of two correlations – Fisher’s Z transformation

Multiple Linear Regression

- ANOVA was extension of T-Test to multiple group means
- Linear regression extends ANOVA to continuous predictor variables
- Multiple linear regression extends simple linear regression to multiple predictor variables
  - Systolic blood pressure predicted by body mass index, sodium intake, and race
  - Body mass index predicted by caloric intake and gender
  - Caloric intake predicted by measure of stress and gender

Reason for Multivariate Regression Models - I

- First, to assess effect of different factors on the subgroup means of an outcome
  - Since factors have differing effects on outcome and are interrelated themselves, necessary to look at a number of predictors simultaneously
    - Systolic blood pressure predicted by body mass index, sodium intake, race
  - We know that each predictor has an independent effect on systolic blood pressure, but how much of the information carried by each one is also carried by the others
    - It could be that 'race' is a surrogate for sodium intake and that sodium intake is actually the cause for high blood pressure
    - Cannot find this out without a multiple regression model

Reasons for Multivariate Regression Models - II

- Second, to predict for an individual the value of the outcome based on the values of the predictor variables
  - e.g., based on body mass index, sodium intake, and race, what would the predicted value of systolic blood pressure for a person
  - Why? To identify those people who may have significantly higher systolic blood pressure than expected from similar people

Multivariate Regression Models - Assumptions

- Predictor variables are ‘fixed’; i.e., same meaning among individuals
- Predictor variable measured ‘without error’
- For each value of the predictor variable, there is a normal distribution of outcomes (subpopulations) with equal variance
- The means of the outcome subpopulations lie on a straight line related to the predictors; i.e., the predictors and the outcomes are linearly related
    - $\mu_{y|x} = \alpha + \beta_1 x_{1j} + \beta_2 x_{2j}$
- The outcomes are independent
- Regression model: $y = \alpha + \beta\ _1 x_{1j} + \beta_2 x_{2j} + \ldots + \beta\ _k x_{kj} + \varepsilon_j$

Interpretation of regression model

- α is the intercept (Y value) when the predictors are zero
- β_k is one of the ‘slopes’ of regression line and represents change in Y for a unit change in X_k with other predictors held constant. All ‘slopes’ interact to produce the overall slope
  - i.e., β_k is the average slope across all subgroups created by the X_k levels
- ε is the error term for each individual and is the residual for that individual
  - Residual is the difference between predicted and observed values

Least squares fit

- Regression parameters ($\alpha + \beta_k$) are determined using method of least squares
- Minimizes the squared differences between each observation and the ‘fitted’ line in the multivariate ‘plane’; i.e., minimizes the residuals
  - Can be dramatically affected by unusual values in the multivariate plane – multivariate ‘outliers’

Approach to developing a regression model

- Determine outcome and plausible predictors
- Plot outcome vs. each predictor to check for linearity
- Fit the regression model and review parameters and tests
- If model has a significant fit and parameters are significantly different from 0, look at residuals to better evaluate fit

Anatomy of Regression ANOVA Table - I

- Coefficient of multiple determination (r^2)
  - Ratio of regression sums of squares (SSR) to the total sums of squares
  - Can be interpreted as the proportion of the total variation in the outcome that can be explained by the regression model
- Total Sums of Squares (SST)
  - SST is total SS in the outcome for the entire sample – analogous to the variance
  - Sum of squared deviations of observed values from the overall mean

Anatomy of Regression ANOVA Table - II

- Error Sums of Squares (SSE)
  - Unexplained variation based on the sum of the squared deviations between the observed and predicted values
  - What is being minimized in least squares

Anatomy of Regression ANOVA Table - III

- Regression Sums of Squares (SSR) – Global fit of the model
  - Determine if any parameters are significantly different from zero
  - Represents variability in the outcome accounted for by regression line
  - Sum of Squared deviations between predicted values and overall mean
  - Part of ANOVA table of most interest
    - It tells if the model explains a significant amount of the variance in the sample
    - If the $\mu_{y|x}$ change little across the range of the x's, they explain very little of the variance in the sample and that predictor (x) will not have a significant F-test in the table

Multivariate and Partial Correlations

- Multivariate correlation is an indicator of strength of the relationship between the outcome and predictor variables
  - Can be tested with an F-test, but F-statistic will be the same as that from the test of the regression sums of squares
- Partial correlation is an indicator of strength and direction of relationship between the outcome and one predictor with the effect of the other variables removed
  - Can be calculated from simple correlations or produced by software

Diagnostic plots

- Residuals vs predicted values
  - Residuals should be scattered randomly around 0 with no discernible pattern
- Predicted vs observed values
  - Should be distributed along a diagonal line from origin to upper right corner

Dummy Variables

- Use ‘dummy’ (indicator) variables to represent the levels of a variable measured on the nominal scale
  - e.g., education, income, race, gender
  - levels have no intrinsic numeric relationship
- Procedure: choose one level as the ‘reference’ level and then form n-1 dummy variables to represent the n levels of the original variable
- Use set of dummy variables in regression model

Example of Constructing Dummy Variables

- Income has three levels: low, middle, and high
- Choose low income as the reference group
- Form two indicator variables:
  - Income1: 0=low, 1=middle
  - Income2: 0=low, 1=high
- Note that both income1 and income2 have to be in the model to properly represent the ‘low’ income group
- A person from the middle income category would have income1=1 and income2 = 0

Interactions

- Interaction terms represent the joint action of two+ variables
- The interaction term of race and income in a model for TV watching is an indicator for whether the relationship of income with TV watching varies significantly by race
- Most effective when constructed using a binary variable (such as race in NGHS) along with another classification variable or a continuous variable
- Mathematically, can construct interactions from any type of variable, but becomes more difficult to interpret

Example of Constructing an Interaction Term

- Determine what variables might represent different relationships
  - Example of race and income from before
- Multiply the two terms together
  - Race*income where race is coded 0=white, 1=black and income is low, middle, high
  - Thus, interaction term applies only to black girls

Interpretation of Interaction Terms

- Test of significance is a test for a different relationship between income and TV watching for black compared to white girls
  - For white girls, relationship of income and TV watching represented by the income ‘main effects’ term
- If test of significance for an interaction term is significant, may be necessary to fit separate models based on the levels of the interaction terms
  - If the race*income interaction is significant, fit separate models for white girls and for black girls

Interpretation of Interaction Terms

- Principal of hierarchy
  - The appropriate main effects must remain in the model if the interaction term remains in the model
  - Otherwise, impossible to interpret with no reference group

Variable Selection

- Start with the full model
- Eliminate most non-significant term and refit the model
- Continue until only significant terms remain
- Keep sets of dummy variables together
- Keep main effects in when interaction terms are in the model even if main effects are not significant
- Step-up and step-wise regression may miss important variables