Applied Statistics 3 Summary

Topics

- Multivariate Analysis
    - Reduction Methods
        - Principle Component Analysis (PCA)
            - Goal
            - Usage
            - Assumptions
            - Determining Coefficients
            - Properties
            - Retaining Components
            - Interpretation
        - Canonical Correlation Analysis
            - Goal
            - Usage
            - Assumptions
            - Intepretation
    - Discriminant Function Analysis
        - Goal
        - Usage
        - Assumptions
        - Determining Coefficients
        - Tests of Significance
    - Classification Analysis
        - Assumptions
        - Usage
        - Error Rates
            - Estimation
        - Nearest Neighbor Classification Rule (KNN)
    - Cluster Analysis
        - Goal
        - Usage

            - Assumptions
            - Types
                - Hierarchical
                    - Linkage Methods
                    - Standardization
                - Partitioning
                    - K-means Clustering
                    - Wards method
                    - K-Medoids
                - Model-based
        - Multi-Dimensional Scaling
            - Classical MD Scaling
                - Goal
        - Correspondence Analysis
            - Interpretation
- Odds & Odds Ratio
    - properties
    - Sampling Distribution of the Log-estimated Odds Ratio
- 2 X 2 Counts
    - Tests of Homogeneity
    - Tests of Independence
    - Sampling Schemes
- Chi-squared Test
    - Mantel-Haenszel
    - R x C
- Fisher's Exact Test
    - Mantel-Haenszel Excess
        - Assumptions
- Generalized Linear Models
    - Types
- Logistic Regression
    - Maximum Likelihood Estimation (MLE)
        - Properties
    - Likelihood Ratio Test / Drop-in-Deviance Test
    - Probit Regression

        - Model Assessment
        - Binomial Responses
            - Model Assessment
            - Extra Binomial Variation (Over dispersion)
                - Determining existence
                - Estimating ψ
        - Multilevel Categorical Responses
            - Ordinal Categorical Responses
- Log-Linear Models (Poisson Regression)
        - Characteristics
        - Interpretation
        - Model Assessment
        - Extra Poisson Variation (Over dispersion)
            - Checking for Over dispersion
- Negative Binomial Regression
- Experiment Design
        - Studies
            - Prospective
            - Retrospective
            - Matched Case-Control studies
        - Research Design Tool Kit
            - Improving Confidence Intervals
            - Choosing a Sample Size
                - Studies comparing 2 proportions
        - Designing a Study
        - Factorial Design
            - 2^2
            - 2^3
            - 2^k

Multivariate Analysis

Reduction Methods

Principle Component Analysis (PCA)

Goal

- Variable and/or Data Reduction
- Create a few linear combos which retain a large amount of the variance

Usage

These principle component combinations can be used in subsequent analysis as explanatory variables.

Assumptions

- Linearity
- Some Correlation among factors

Properties

Let a given principle component be represented by z_j

- Ordered by variance of Z
- Expect Most information to be contained in the first few components
- $[z_1, z_q]$ are uncorrelated
- $\sum var(z_j) = \sum var(y_j)$

Determining Coefficients

- $Var(Z_1)$ maximized with constraint $a_1' a_1 = 1$
- $Var(Z_2)$ maximized with constraint $a_2' a_2 = 1$ and $cov(Z_1, Z_2) = 0$.

To generate coefficients, use:

- The original variables' covariance matrix (if using original vars)
- The original variables' correlation matrix (if using standardized original vars)

Original vs Standardized Vars

Orignal

- Easier to interpret
- Results dependent on unit of measurement
- Principle Components tend to reflect vars with largest variance

Standardized

- Can be used when vars are of difference scales
- More difficult to interpret
- More common

Retaining Components

Two options

1. Enough should be retained to explain 80% of the total variation
2. Lower bound on the number of retained components
    - $\frac{\sum lambda_i}{q}$ (average eigenvalue)
        - where $lambda_i$ are the eigenvalues of the covariance matrix (original) or correlation matrix (standard)
    - avg eignevalue = 1 when using the correlation matrix
3. Scree plot
    - Plots eigenvalues vs the component #
    - Choose # of components where the plot begins to *flatten* out

Interpretation

Each Principle Component describes how a group of variables are interrelated.

- Focus on *loadings* (coefficients). Loading > 0.5 helps determine which vars are influential

- If all elements of the first eigenvector/coefficients/loadings are positive, Principle Component measures *size*
- If some are positive and negative, then the PC measures a difference of the variables
- If all loadings are roughly the same size in magnitude and the same sign, the PC can be interpreted as a weighted average

Canonical Correlation Analysis

Start with two sets of variables:

1. A set of response variables
    - y1, y2, ..., yq
2. A set of explanatory variables
    - x1, x2, ..., xq

Goal

$$U = \sum_{i=1}^{q} a_i y_i$$

$$V = \sum_{i=1}^{p} b_i x_i$$

Find coefficients a1, a2, ..., aq and b1 , b2, ..., bq that maximize the correlation between U and V.

These are called canonical variates which are essentially linear combinations of the original two sets of variables.

Number of Canonical correlations: $s = min(p, q)$

- U_i and V_i are correlated. i = [1, s]
- U_i and V_j are uncorrelated. i $=\!\!\!\!/$ j
- V_i and V_j are uncorrelated. i $=\!\!\!\!/$ j
- U_i and U_j are uncorrelated. i $=\!\!\!\!/$ j

Canonical Correlation Sqaured: $[r_{\underset{1}{2}}, r_{\underset{s}{2}}]$

- Proportion of variance explained in the dependent vars (Y's) explained by the independent set of vars (X's) along a given dimension (s dimentions)

Redundancy Analysis: Explains variation by evaluating the adequacy of prediction from the canonical analysis,

Usage

- Measure correlation between X's and Y's
- Extension of multiple correlation ($\sqrt{R_2}$)
- Often a compliment to multivariate regression

When to use

- Regression analysis appropriate but more than one dependent variable Y
- Useful when dependent variables are moderately inter-related
- Can be used to test independence between the independent vars (X's) and dependent vars (Y's)

Assumptions

- Linearity of Correlations
- Linearity of Relationships
- Multivariate Normality
    - Desirable since it standardizes a distribution to allow for a higher correlation among variables
    - Highly recommended that all vars are evaluated for normality and transformed if needed

Intepretation

Low p-values indicate significance of a correlation. In an example with 4 canonical correlations:

- CV_1: $H_0 : \rho_1 = \rho_2 = \rho_3 = \rho_4 = 0$

- CV_2: $H_0 : \rho_2 = \rho_3 = \rho_4 = 0$
- CV_3: $H_0 : \rho_3 = \rho_4 = 0$
- CV_4: $H_0 : \rho_4 = 0$

Discriminant Function Analysis (DFA)

Goal

Classify a subject or unit into two or more groups based on info collected on independent variables. Groups **must** be clearly defined.

How likely is a subject in $group_j$ based on the basis of a set of quantitative variables?

Usage

Come up with a single set of coefficients to apply to all groups **then** Construct linear combinations of these variables and use them to distinguish populations.

Distribution between groups?

- Yes: parametric methods (linear or quadratic DFA)
- No: non-parametric method

Assumptions

- Equal Spread
- Some Assume Normality

Determining Coefficients

- Maximize separation between two groups
  Mahalanobis distance

$$D_2 = \frac{(\bar{z}_1 - \bar{z}_2)_2}{s_{z}^{2}}$$

Multi-dimensional generalization of measuring how many std devs away from a point is the mean (or centroid) of the distribution (Like a Z-score).

Scalings from LDF are not the same as $A_T = S_{pl}^{-1}(\bar{y}_1 - \bar{y}_2)$

K Groups

Goal

Find a vector A that maximizes separation between $[\bar{z}_1, \bar{z}_k]$

Usage

How?

- Replace $(\bar{y}_1 - \bar{y}_2)_T$ with the H matrix from MANOVA
- Replace S_{pl} with E matrix

H indicates spread between groups
E indicates spread within each group

$$\lambda = \frac{a_T H a}{a_T E a}$$

$$\rightarrow a_T(Ha - \lambda Ea) = 0$$

$$\rightarrow (E_{-1}H - \lambda I)a = 0$$

Solutions are the eigenvalues $[\lambda_1, \lambda_s]$ and eigenvectors $[a_1, a_s]$ of $E_{-1}H$ where $s = rank(H) = min(k-1, s)$

From eigenvectors $[a_1, a_s]$ of $E_{-1}H$, s **discriminant functions** are obtained:

- $z_1 = a_T y$
- $z_2 = a_T y$

  ...
- $z_s = a_T y$

These discriminant functions are uncorrelated. They show the dimensions or directions of differences among $[y_1, y_k]$. The relative importance of each discriminant function can be assessed by considering its eigenvalue as a proportion of the total.

$$\frac{\lambda_i}{\sum_{j=1}^{s} \lambda_j}$$

Matrix $E^{-1}H$ is not symmetric. Special computation must be done in R:

- Find matrix U that is the Cholesky factorization of E. $E = U^T U$
- Find the eigenvector b of the matrix $(U^{-1})^T H U^{-1}$
- $a = U^{-1} b$ is an eigenvector of $E^{-1}H$

Tests of Significance

- Two Group Case
  - Use Hotellings T^2 to test $H_0 : a = 0$
- K Group Case
  - Wilks' lambda since eigenvalues are the same as eigenvalues from MANOVA
  - $V_m = [N - 1 - \frac{1}{2}(p - k)] \sum_{i=0}^{k} log(1 + \lambda_i)$
    - p = # of vars
    - k = # of groups
    - $V_m \approx \chi^2$ (p - m + 1)(k - m) degrees of freedom

Forward, Backward, or Stepwise Selection can be performed to determine predictors that are most significant for discriminating against others.

Interpretation

Standardizing helps. The largest magnitude contributes most to the equation (similar to PCA and CCA).

Classification Analysis

The predictive aspect of Discriminant Analysis. Synonyms include Discriminant Analysis, Pattern

Recognition, and Cluster Analysis.

Assumptions

- No assumptions around distributions
- $\Sigma_1 = \Sigma_2$ (Equal covariance matrices)
    - If violated, Distance function is $D_i^2(y) = (y - \bar{y}_i)_T S_i^{-1} (y - \bar{y}_i)$ where S_i is the sample covariance for the i_{th} group.

Usage

1. Obtain a sample of observation vectors from each group
2. Choose a Sampling Unit whose group membership is unknown
3. Assign unit to a group based on vector of p measure values (y) associated with the unit*

* If prior probabilities p_1 and p_2 are known for 2 populations, the classification rule can be modified.

Normal Base Classification Rule

- $f(y|G_1) \approx N_p(\mu_1, \sigma)$
- $f(y|G_2) \approx N_p(\mu_2, \sigma)$

Error Rates

Error Rate: probability of misclassification
Correct Classification Rate: Complement of Error Rate

Estimation

A simple method is to plug the values back in and see how many matched. For large samples, the error rate has a small amount of bias for estimating the actual error rate.

For small samples, **Holdout/leave-one-out/Cross Validation**. All but one observations used to compute the classification rule then used to classify the omitted observation

Nearest Neighbor Classification Rule (KNN)

Compute distance from y_i to all other points y_i using distance function

$$(y_i - y_j)_T S_{\underset{pl}{-1}}(y_i - y_j), j \neq i$$

If a majority of K points belong to G_1, assign y_i to G_1, else G_2

Choosing K

- $K = \sqrt{n_i}$
- Try several values of k and use the one with the best error rate

Cluster Analysis

Goal

Separate Individual observations/items into groups/clusters on the basis of values for p variables measured on each variable

- items/objects == rows
- Distance measured is typically Euclidean

Type of unsupervised classification because the nature or number of groups is not necessarily known prior to classification

Usage

Assumptions

- N objects/cases/rows of data
- K clusters/groups
    - If K is known, the number of ways to partition N into K is a **stirling number of the second kind**
    - If K is not known, the number of possible partitions is much larger

Types

Hierarchical

Cluster data in a series of n steps, joining observations together step-by-step to form clusters.

- Fast computation for small datasets
- Dendogram for visualizing a variety of k clusters

Linkage Methods

Single Linkage or Nearest Neighbor

Join clusters whose min distance between object is smallest

$$D_{AB} = min(d_{ij})$$

where d_{ij} is the distance between an element in A and B.

Complete Linkage: Single Linkage with max instead of min.
Average Linkage: Single Linkage with avg instead of min.

Standardization

- Divide each column by its sample std dev so all variables have a std dev of 1
- Divide each variable by its sample range
- Z Scores

Partitioning

For a fixed value of K, seek the best possible partition for that K which optimizes some objective function.

K-means Clustering

Find the partition of N objects into K clusters that minimize within-cluster SS. Traditionally, distance between clusters is euclidean. Goals is to minimize the sum of squared euclidean distances

$$WSS = \sum_{c=0}^{k} \sum_{i=0}^{n} d^2_{E(y_i, \bar{y}_c)}$$

Final clustering result dependent on initial configuration of rows. Good to rerun the algorithm a few times with different starting points to ensure stable results.

Wards method

Mix of Hierarchical and K-means. Each object starts as its own cluster and concludes with all objects in one cluster. At each step, the method searches all possible ways to join a pair of clusters so that the WSS is minimized for that step.

K-Medoids

Robust alternative to K-Means. Minimizes

$$C_{md} = \sum_{c=0}^{k} \sum_{i=0}^{n} d(y_i, m_c)$$

M_c is a medoid (most representative object). Best to think of it as a p-variate median. Like K-Means, K-Medoids does not globally minimize its criterion in general.

Pros

- Accepts a dissimilarity matrix as well as raw data matrix
- Generates silhouttes for K-clusters so don't need to decide K ahead of time

Cons

- Computationally infeasible for n > 5000.

Other criteria for choosing k include the Dunn Index and the Davies-Bouldin Index

Model-based

Assumes an underlying distribution for the K clusters.

Multi-Dimensional Scaling

Use distances to measure how different multivariate observations were from each other. Can take a multivariate dataset (a set of p-dimensional vectors) and calculate distances between pairs of vectors.

Classical MD Scaling

Goal

Given an N x N matrix, construct a map containing multivariate points. There are no unique or best solutions where to place points on map.

Sometimes referred to as Principle Coordinates Analysis.

Correspondence Analysis

Contingency Table presents sample values for two categorical variables and test for independence between the two. This supplements a chi-square test

Chi-square distance: Column Proportions with entries $p_{ij} =$

$$\frac{n_{ij}}{n_i}$$

Interpretation

With all rows and categories plotted:

- Two row categories near each other have similar conditional distributions across columns
- Two column categories have similar profiles
- A Row and Column Category near tend to appear more ofthen than expected under independence.

Odds & Odds Ratio

The probability of something happening (ω)

Properties

- $\omega \geq 0$
- If P (probability) = 0.5, them $\omega = 1$ (50-50 odds)
- If ω is odds of success,

$$\frac{1}{\omega}$$

is odds of failure

- $P =$

$$\frac{\omega}{\omega+1}$$

Odds Ratio

Example

$$\phi = \frac{\omega_1}{\omega_2}$$

$$\rightarrow \omega_1 = 5\omega_2$$

The odds of "success" in Group 1 is 5 times the odds of "success" in Group 2

$$\phi = \frac{\omega_1}{\omega_2}$$

	Response
	Yes	No
1	n_{11}	n_{12}
2	n_{21}	n_{22}

$$\phi = \frac{n_{11}n_{22}}{n_{21}n_{12}}$$

Odds Ratios vs. Population Proportion

- ϕ tends to remain more nearly constant over levels of confounding variables
- ϕ is the only parameter that can be used to compare groups of responses from a **retrospective** study
- ϕ extends into Logistic Regression models

Sampling Distribution of the Log-estimated Odds Ratio

- $E(\log(\hat{\phi})) \approx \log(\phi)$
- $Var(\log(\hat{\phi})) \approx$

$$\frac{1}{n_1 p_1 (1-p_1)} + \frac{1}{n_2 p_2 (1-p_2)}$$

    - Similar to a binary distribution
- if n_1 and n_2 are sufficiently large, the sampling distribution is approximately normal

2 X 2 Counts

$H_0 : \pi_1 - \pi_2 = 0 \equiv H_0 : \frac{\omega_1}{\omega_2}$

Tests of Homogeneity

Is a binary response the same across multiple populations?

Tests of Independence

Is there an association between row and column factors without specifying one of them as a response variable? Refers to a single population.

H_0: The row category is independent of the column category

Sampling Schemes

Odds Ratio can be used with any Sampling Scheme

Poisson

Frequency of success over a period of time or space. Random sample from a single population where each member falls into a cell in an R x C table.

No Marginal Totals known in advance

Used for tests of **homogeneity** and **independence**

Multinomial

K categories for a sample of N. Similar to Poisson except **Total Sample Size (T) is fixed in advance**.

Used for tests of **homogeneity** and **independence**

Prospective Product Binomial

More than one Binomial Distribution is present. Random samples selected from each population

Row Totals fixed in advance

Used for Test of **homogeneity** but only for the odds ratio

Retrospective Product Binomial

Flip explanatory and Response variable from Prospective Binomial Sampling

Column totals fixed in advance

Used for Test of **homogeneity** but only for the odds ratio

Randomized Binomial Experiment

Subjects randomly allocated to the two levels of the explanatory factor (Rows of the table). This follows Prospective Product Binomial except instead of random sampling, randomization of subjects into groups is used.

Used for Tests of **homogeneity**

Hypergeometric Probability Distribution

If interest is stricly focused on the odds ratio, analysis may be conducted conditionally on the row and column totals

Both row and column totals are fixed

Used in Fisher's Exact Test

Chi-square Tests

Pearson Chi-Square Test for Goodness of Fit

Determine GoF based on the assumption that the expected count follows a χ_2 distribution.

Observed Count: Number of units that fall into a cell.
Expected Count: Number of units predicted by theory to fall into a cell when H_0 is true

$$\chi_2 \approx \sum \frac{(Observed - Expected)_2}{Expected}$$

If H_0 is true, then the chi-square test approximates χ_2 with df = number of cells - 1

Chi-Squared Test of Independence in an R X C Table

When H0 is true, sampling distribution of χ_2 has an approximate χ_2 distribution with (r - 1) X (c - 1) df where r is the number of rows and c is number of columns.

Limitations

- Only Product is a p-value
- No associated parameter to describe the degree of dependence

        - look at expected ratios vs actual ratios to determine dependency
    - Alternative Hypothesis very general
    - When 3+ rows and columns involved, may be a more specific form of dependence to explore

Mantel-Haenszel

A more powerful alternative to the Pearson Chi-square Test when at least one of the factors are **ordinal**. An ordinal may be defined as a midpoint for a range of response variables.

r = some measure of the sample correlation between two factors
n = sample value

$$M_2 = (n-1)r_2$$

$$H_0 : \rho = 1$$
$$H_A : \rho \neq 1$$

Sampling Distribution of $M_2 \approx \chi_2$ with df = 1 under H_0

Fisher's Exact Test

Randomization test based on statistic $\pi_1 - \pi_2$. When data is observational, it can be thought of as a permutation test. This is a useful interpretation when the entire population has been sampled or a sample is not random.

- Inference possible for Poisson, Multinomial, and Product Binomial sampling schemes

Can be used for tests of equal population proportions, equal population odds, or independence

Mantel-Haenszel Excess

Excess: Observed Count - expected count in one cell of a R x C table. This is like a residual for

cell counts.

Excess of n_{11} = n_{11} $-$

$$\frac{R_1 C_1}{T}$$

Under H_0:

- E(Excess) = 0
- Var(Excess) =

$$\frac{R_1 R_2 C_1 C_2}{T T (T-1)}$$

For a 2 x 2 table of counts, excess is an approximation of Fisher's Exact Test.

An overall association can be developed for a third factor. A weighted average of the odds ratios across 2 x 2 tables should be calculated. This treats the third factor as a block.

Tests for conditional independence and homogenous assocation for the k conditional odds ratios in K 2 x 2 tables. It combines sample odds ratios for the partial K tables into a single summary measure of partial assocation.

Appropriate for prospective, retrospective observational data, and randomized experiments.

Assumptions

- Odds Ratio same in each 2 x 2 Table. (Use Breslow-Day Statistic)
    - H_0: X and Y are conditionally independent given Z ($\theta_{XY(k)} = 1$)
- Sum of expected counts over all tables should be at least 5.

Generalized Linear Models

Probability Model in which the mean of a response variable is related to explanatory variables through a regression equation. There is a function out there which converts a response variable to a linear function. This is called the **link function**.

Types

Link Function: A specified function of μ equal to the regression structure. The non-linearity is contained within the link function.

$g(\mu) = \beta_0 + \sum_{i=1}^{p} \beta_i X_i$

Normal

Used for Ordinary Least Squares (OLS) Regression
Link: Identity
Function: $g(\mu) = \mu$

Poisson

Used to count occurrences in a fixed time or space
Link: Log
Function: $\log(\mu)$

Bernoilli, Binomial, Categorical, Multinomial (Logistic)

Outcome of a single binary response OR
number of successes OR
outcome of a single K-way occurrence
Link: Logit
Function:
$\log(\frac{\pi}{1-\pi}$
$\log(\theta) = \beta_0 + \sum_{i=1}^{p} \beta_i X_i = \eta$

Known as the log-odds because it is a log function of the odds where the odds of success = π

Logistic Regression

Logistic Function: Inverse of the Logit function

$$\pi = \frac{e_{\eta}}{1 + e_{\eta}}$$

- $E(Y) = \pi$
- $Var(Y) = \pi(1 - \pi)$

$$\omega = \frac{\pi}{1 - \pi}$$

$$\omega_A = e_{\beta_0 + \beta_1 A}$$
$$\omega_B = e_{\beta_0 + \beta_1 B}$$

Odds of A/B

$$\omega_{AB} = \frac{\omega_A}{\omega_B}$$

Maximum Likelihood Estimation (MLE)

$$Pr(Y = y) = \pi_y (1 - \pi)_{1-y}$$

Joint Probability Mass Function

$$P(Y_1 = y_1, ...) = \prod Pr(Y_i = y_i)$$

To find MLEs of the Logistic Regression coefficients, set each p + 1 partial derivatives to 0.

Solutions for parameters to this system of equations are MLEs for the LR coefficients. The solution to this system does not exist in closed form; therefore, iterational computational procedures such as the Newton-Raphson, are used.

Properties

If a model is correct and the sample size is large enough:

- MLEs are essentially unbiased
- Formulas exist for estimating the std devs of the Sampling Distribution of the

Estimators
- Estimators are MVUE
- The sampling distribution is approximately Normal

When working with Asymptotic Normal Results, these procedures are called Wald procedures. They assume large sample sizes make everything statisically okay.

Likelihood Ratio Test / Drop-in-Deviance Test

Analogous to Extra Sum of Squares F-Test in Linear Regression. Compares a full model to a reduced model. When H_0 is true, the reduced model is the correct model.

$LRT \approx \chi_2(\nu)$ where ν = diff(num_param_full, num_param_reduced). With GLMs, a quantity called Deviance is used.

LMAX = Maximum Likelihood Function

Deviance = Sum of Squared Residuals = -2 * log(LMAX)

$LRT = 2\log(LMAX_{full}) - 2log(LMAX_{reduced})$ = Deviance_full - Deviance_reduced

To test significance of a single term, DinD test between full model and full model minus the single term. This is not the same as Wald's test for a single coefficient. If the two give different results, DinD has a more reliable p-value.

Model Assessment

- For model terms, Informal testing of extra terms such as squared or interaction terms is important.
- For model adequacy
    - Hosmer-Lemeshow GoF Test
    - Deviance Residual Plots vs Predicted Values and each of the predictor variables
        - Loess function should be as flat as possible
    - More complicated GoF tests exist

AIC / BIC

AIC = deviance + log(n) * p
BIC = deviance + 2 * p

Probit Regression

Any cumulative distribution function $F(\pi)$ has characteristics similar to the logit function. Typically, $F(\pi)$ is chosen to be the inverse of the Normal CDF. As long as $\pi : [0.2, 0.8]$, it is similar to logistic regression.

Binomial Responses

$Y_i \approx Bin(m_i, \pi_i)$

Model Assessment

- Scatterplots: Empirical logits vs Explanatory Variables
    - log odds vs explanatory vars. Log-odds on Y, explanatory on X. If it looks linear, Good! Otherwise, a transformation may be needed
- Examining Residuals
- Deviance GoF test
    - DinD with intercept-only-model and proposed model

Examining Residuals

Deviance Residual: sum of all n squared deviance residuals = deviance statistic. This measures discrepency in the likelihood function to the fit of the model at each observation.

Pearson Residual: Observed Binomial Response variable minus estimated mean, divided by estimated std dev. (like Z Score). Roughly mean = 0 and var = 1.

With at least 5 trials in any binomial response, then any residual greater than 2 in magnitude may be a possible outlier. No discernable pattern indicates the error terms are normal

Extra Binomial Variation (Over dispersion)

If binomial trials are not independent or important explanatory variiables are not included in the

model for π_i, response counts will no longer have binomial distributions

When Over dispersion is present, regression parameter estimates will not be seriously biased but standard errors tend to be smaller leading to small p-values, narrow C.I., and mistaken interpretations.

The **quasi-likelihood approach** assumes a relationship between mean and Var(Y) rather than a specific probability distribution for Y. The variance formula is multiplied by an estimated constant ψ. $\psi > 1$ indicates overdispersion.

It does not affect regression coefficients but it affects standard errors.

Determining existence

Yes to any of these questions cautions the use of the binomial model.

- Are binary responses included in each count unlikely to be independent?
- Are Observations with identical values in explanatory variables likely to have different π_i's?
- Is the model for π naive?

Estimating ψ

$$\hat{\psi} = \frac{\sum Pres_2}{n - p\ i}$$

Let D = number of parameters in the full model.

$$F = \frac{\frac{DinD}{D}}{\hat{\psi}}$$

Multilevel Categorical Responses

Let J be the number of categories for Y and $[\pi_1, \pi_i]$ denote the reponse probabilities.

Multicategory logit models simultaneously use all pairs of categories by specifying the odds of

outcome in one category instead of another.

$$\log(\frac{\pi_a}{\pi_b}) = (\beta_{0a} + \beta_{1a}x) - (\beta_{0b} + \beta_{1b}x) = (\beta_{0a} - \beta_{0b}) + (\beta_{1a} - \beta_{1b})x$$

Ordinal Categorical Responses

When response categories are ordered, logits can utilize the ordering. Using a cumulative logit function, the outcome for is the probability that a value Y falls below a category J. It looks like a binary logistic regression model.

$$logit[PrY \leq j] = \log[\frac{PrY \leq j}{1 - Pr(Y \leq j)}] = \log[\frac{\sum_{j}^{i=1} \pi_i}{\sum_{J}^{l=j+1} \pi_l}] = \beta_{0j} + \beta_1 x$$

Log-Linear Models (Poisson Regression)

For Y, the number of successes in a given time or space interval. The Poisson Distribution is most appropriate for counts of rare events that occur at completely random points in space or time. Works reasonably well for count data where spread increases with mean.

Characteristics

- $Var(Y) = \mu(Y) = \mu$

- Distribution tends to be right-skewed and is most pronounced when the mean is small
- Larger means tend to be well approximated by a normal distribution

Log link helps straighten the relationship between the predictors and the response; however, variance will still be non-constant after the transformation.

Interpretation

Multiplicative effect on the mean. Can also convert to an estimated percent increase. ($e_{\beta_1} = proportion$). This is different than logistic regression where e_{β_1} gives the odds ratio.

Model Assessment

- Scatterplots
- Residuals
    - Deviance Residual (more reliable for detecting outliers)
    - Pearson Residual
- Deviance GoF

If Poisson means are at least 5 (large):

- Distribution of both residuals are approximately Standard Normal
- If > 5% of residuals exceed 2 in magnitude or if one or two greatly exceed 2, there are problems in the fit

If Poisson means < 5 (small):

- Neither set of residuals follows a normal distribution well thus comparison to standard normal provides a poor lack of fit

Deviance GoF Test: informal assessment of the adequacy of a fitted model

- Use in conjunction with plots and tests of model terms
- Large p-value indicates model is inadequate OR insufficient data to detect inadequacies

- Small p-value indicates Model for the mean is incorrect OR Poisson is an inadequate model for the response OR a few severely outlying observations contaminate the data

Extra Poisson Variation (Over dispersion)

Over disperson leads to higher variance in responses than predicted by the Poisson Distribution.

- Unmeasured effects
- Clustering of events
- Other contaminating influences

Checking for Over dispersion

- Is it likely?
    - Are important explanatory variables not available?
    - Are individuals with the same level of explanatory variables behaving differently?
    - Are events making up the count clustered or systematically space rather than randomly spaced?
- Fit a negative binomial model and check if $\psi > 1$
- Compare Sample Variance to Sample Averages for groups of responses with identical explanatory variable values
- Examine Deviance GoF Test after fitting a rich model
- Examine Residuals to see if a large deviance statistic may be due to outliers

Negative Binomial Regression

Alternative to quasi-likelihood estimation in Poisson regression when over dispersion is present is negative binomial regression. An additional parameter ϕ is used to model count variation.

- $\mu Y_i|X_{i1}, ..., X_{ip} = \mu_i$
- $Var(Y_i|X_{i1}, ..., X_{ip}) = \mu_i(1 + \phi\mu_i)$

Strategies same as log-linear regression except no need to investigate extra-poisson variation.

Experiment Design

Studies

Prospective

Subjects selected from or assigned to group with specified explanatory factor levels then responses are determined. This is the traditional experiment design.

Retrospective

Subjects selected from groups with specified response levels then their explanatory factors are determined. Only the odds ratio (ϕ) can be estimated.

This is useful if response proportions are small which would normally require huge samples in a prospective study (i.e. Cancer Rates). It is also useful if there would be moral implications for conducting an experiment in a prospective fashion (i.e. link between smoking and cancer).

Matched Case-Control

In a 2 x 2 table, case-control studies match a single control with each case. For a binary response Y, each case (Y = 1) is matched with a control (Y = 0) according to a certain criteria that could affect the response. The study observes cases and controls on the predictor variable X and analyzes the XY association.

Analysis uses **Conditional Likelihood Logistic Regression**. Each subject has their own probability distribution.

$$\log\left(\frac{\pi_{i1}}{1-\pi_{i1}}\right)$$

$$\log\left(\frac{\pi_{i2}}{1-\pi_{i2}}\right)$$

β_{0i} allows probabilities to vary among subjects. It can be extended to K predictors but typically one variable is of special interest while the others are controlled covariates.

Research Design Tool Kit

Controls and Placebos

- Control provides baseline
- Placebo mimics new treatment in every aspect except the test ingredient

Blinding

- Subjects do not know which treatment is being received. This eliminates the possibility that the end comparison measures expectations rather than results

Blocking

- Arrange units into homogenous subgroups in which treatments are randomly assigned to units in each block
- Strives to improve precision, control for confounding variables, and expand scope of inference about treatment differences

Stratification

- Population units partitioned into homogenous subgroups (strata) and a random sample from each stratum is obtained

Covariates

- Auxilary measurements taken on each unit
- Doesn't directly address the question but may be closely related
- Controls for potentially confounding factors, improves precision, assess the model, and expands scope of inference

Randomization

- Random Procedure to assign experimental units to different treatment groups
- Controls for factors not explicitly controlled for in the design or the analysi
- Permits Causal inferences
- provides a probability model for drawing inferences

Random Sampling

- Employ a Well-Understood random procedure to select units from a population

Replication

- Conducting copies of a basic study pattern
- Refers to assigning one treatment to multiple units with each block
- increased precision for treatment effects and improved model assessment

Balance

- Having the same number of units assigned to each treatment group
- Optimize precision for treatment comparisons and ensure independence

Improving Confidence Intervals

$$\bar{Y}_1 - \bar{Y}_2 \pm qt(1-\alpha/2, n_1 + n_2 - 2)s_p \sqrt{\quad}$$

- qt reduced with replication or α
- s_p reduced by blocking, including covariates, or improving measurement technology
- square root term can be reduced by replication or balance

Choosing a Sample Size

$$n = 4 \frac{[qt(1-\alpha/2, n-k)]_2 s_2}{(PracticallySignificantDifference)_2}$$

where C_k is the kth coefficient in a linear combination of means g.

Studies comparing 2 proportions

Designing a Study

1. State objective
2. Determine Scope of Interest
    - Will it be randomized or observational?

        - What experimental or sampling units will be used?
        - What are the populations of interest?
3. Understand the system under study
4. Decide how to measure the response
5. List factors that can affect the response
        - **Design factors**: factors to vary; factors to fix
        - **Confounding factors**: factors to control (blocking); factors to control by analysis (covariates); factors to control by randomization
6. Plan the conduct of the experiment (timeline)
7. Outline the statistical analysis
8. Determine Sample Size

Factorial Design

2^2

2^3

2^k

Coded variable form (-, +) useful for experimenter

- Gives all effects and interactions
- T-statistics equivalent to F-statistics

Engineering Units form useful for others

- Does not depend on experimental levels or factors
- Coefficients have a different interpretation: a regression coefficient represent the effect of changing a factor by 1 (engineering) unit, not the effect of changing from low to high