Generalized Linear Model

Normal	$g(m) = m$	$m = x^T\beta$
Poisson	$g(m) = log(m)$	$m = e^{x^T\beta}$
Bernoulli	$g(m) = log(\frac{m}{1-m})$	$m = \frac{e^{x^T\beta}}{1+e^{x^T\beta}}$
Gamma	$g(m) = \frac{1}{m}$	$m = \frac{1}{x^T\beta}$

- Poisson Distribution (i.e., rate)
  - Variance of response = expectation
  - Variance not assumed to be constant
  - Standard LR w/ log transformation causes violations in constant variance
- Exponential Distribution (i.e., wait time)
- Other Distributions
  - Gamma, Bernoulli (Binomial)

Poisson Regression (using Maximum Liklihood Estimation to estimate model parameters)

- log function is the log rate $ln(\lambda(x)) = \beta + \beta_1 x$
- with an increase wiht one unit in x (if quantitative): $\frac{e^{\beta_0+\beta_1(x+1))}}{e^{\beta_0+\beta_1 x}} = e^{\beta_1}$
- if categorical with respect to the baseline: $\frac{e^{\beta_0+\beta_1(x=1))}}{e^{\beta_0+\beta_1(x=0))}} = e^{\beta_1}$
- interpret regression coefficients in terms of log ratio of the rate keeping all other var constant
- No error term: p+1 regression coefficients if including intercept (p = # of predict. variables)
- Constant variance does not hold for poisson regression

Example:

- Test using standard LR -> Test to see if variance of residuals is constant therefore use Poiss
- For one unit increase, the log expected [response] increases by XXX, holding other var fixed
- The rate ratio for [response] would be expected increase by a factor of exp(XXX)=ANS

Statistical Inference:

- MLE assumption of normal relies on the assumption of lage sampel size -> not reliable for small sample data
- Use Z-test (Wald test) for statistical significance of parameter -> normal distribution (not t as in standard regression)
- Small sample sizes causes more type 1 errors than expected
- Testing for Subsets of Coefficients
  - Null Hypothesis: All alpha coefficinets (those not in reduced model) = 0
  - Alternative: At least of the parameters not included does not equal 0
  - Use wald test (Terms argument is the terms that need to be tested)
- Overall Regression
  - Similar but use difference in deviance between full and null models. DOF = # of variables
  - Use chi-squared distribution (1 - pchisq((null dev - resid dev), (null DOF - resid DOF))
    - Small p-value reject null hypothesis and determine that at least one predicting varaible significantly explains the variability

<u>confidence int.</u>

$$\hat{\beta}_j \pm z_{\frac{\alpha}{2}}\sqrt{V(\hat{\beta}_j)}$$

Goodness of Fit:

- Poisson regression assumptions (No error terms!)
  - Linearity Assumption $log(E(Y|x_1, ..., x_p)) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$
  - Independence Assumption : Y1, ..., Yn are independent random variables
  - Variance Assumption: $E(Y|x_1, ..., x_p) = V(Y|x_1, ..., x_p)$
    - Don't need to assume the variance is constant
- Pearson residuals follow directly a normal approximation to a binomial
- How to evaluate?
  - Use person residuals to identify if they are normally distributed (if normal then good fit)
  - Hypothesis testing (want large p values to fail to reject null hypothesis)
    - Null hypothesis: Poisson model fits the data (chi-squared dist = n-p-1 DOF)
    - Alternative hypothesis: Poisson model does not fit the data
- Not a good fit -> what to do?
  - Add predicting variables, transform predicting variables to imporve linearity, inter. terms
  - Identify outliers
  - Poisson distribution isn't appropriate:
    - **Overdispersion**: Variability of the est. rates is larger than implied by Poisson model
      - Correlation in observed responses, heterogeneity in rates that hasn't been modeled
      - $\hat{\phi} = \frac{D}{n-p-1}$ where D is the sum of the squared deviances, \phi > 2 then overdisp.
- Example:
  - p = 1-(pchisql(resid deviance, resid DOF)) -> if greater than alpha then good fit
  - Still need to test for residual normality

Goodness of Fit (Logistic Regression)

- Logistic regression without replications (ni = 1)
  - One seperate set of predictors for each observation (i.e., each observation is an independent trial)
- Logisitic regression with replications (ni > 1)
  - Binomial distribution of repeated responses for each unique set of predictors
- Residuals can only be defined **with** replications
  - Only perform goodness of fit **with** replications
- Ways to test
  - Visual GOF Interpretation
    - Normal probability plot and normal distribution in histogram of residuals
  - Hypothesis testing

    H_0: Logistic regression model fits the data

    H_a: Logistic regression model does not fit the data
    - Want large p-values so we fail to reject the null hypothesis
  - Goodness of fit -> model assumptions hold (1-pchisq(deviance([model]), DOF))
    - Can also use pearson residuals
    - DOF is residual DOF
- Simpsons paradox: Sign of coefficient changes depending on marginal/conditional

Variable Selection

- Objectives
  - High Dimensionality
  - Multicollinearity
  - Predictive vs Explanatory Objective
    - Explanatory purpose: Can use correlated predictors
  - Meet Bias vs Variance Tradeoff
    - Many # of predictors -> Low Bias, High Variance
    - Low # of predictors -> High Bias, Low Variance
- Training Risk: Biased estimator of prediction risk
- Criteria:
  - Mallows Cp
  - AIC
    - k = 2
    - Measure of prediction risk
  - BIC
    - Penalizes complexity more than other approaches
    - k = log(n)
- Methods
  - Stepwise
    - Backward: More computationally intensive than forward
    - Not a greedy algorithm: Does not check all possible combinations
    - Forward regression preferred -> Starts with smaller model
  - Ridge Regression (Regularized Regression)
    - Reduces standard error by adding bias to estimators
    - Minimizes SSE plus L2 error term
  - LASSO Regression (Regularized Regression)
    - Does not perform well under multicollinearity
    - Requires algorithm to minimize the penalized sum of least squares
    - Minimize sum of absolute value of regression coefficients
  - Elastic Net (Regularized Regression)
    - Combines variable selection of LASSO and minimization of standard error of ridge (i.e. uses both penalties of ridge and LASSO)
- Penalties
  - L0: Provides best model given criteria while searching through all models
  - L1 (LASSO): Measures sparsity and forces coefficients to 0
  - L2 (Ridge): Does not do variable selection
- Notes
  - Variables that are used to control bias should be forced into the model
  - Penalization in linear regression means penalizing for for complex models
  - Can be applied even when number of predictors is larger than observations
  - Must standardize/rescale varaibles before regularized regression