Epidemiology/Public Health 576C
Review of Statistical Modeling

Overview

	Linear	Logistic	Survival	Poisson
Outcome	Continuous, normal	Binary	Time-to-event with censoring	Count
Model	$E(y) = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \ldots + \beta_k \cdot x_k$	$\ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \ldots + \beta_k \cdot x_k$	$h(t; x_1, x_2, \ldots x_k) = h_0(t) \cdot e^{\beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \ldots + \beta_k \cdot x_k}$	$\log(q_{ij}) = \alpha + \beta_i + \delta_j$
Parameter Estimation	Least squares = maximum likelihood (assuming normality)	Maximum likelihood	Maximum likelihood	Maximum likelihood
Hypothesis Testing – single variable	*t* test	Wald statistic	Wald statistic	Wald statistic
Hypothesis Testing – multiple variables	F test	Likelihood ratio test	Likelihood ratio test	Likelihood ratio test
Diagnostics for assumptions	Plots for linearity Normality of residuals Constant variance	Linear in log-odds	Proportional hazards	No over- or under-dispersion
Influential observations	Cook's Distance	Pregibon's Delta-Beta	Likelihood displacement value	
Goodness-of-fit	R^2	Area under ROC curve Hosmer-Lemeshow Goodness-of-fit test		Goodness-of-fit test

Modeling Strategies

The more appropriate variable selection method depends on the purpose for the model. Three distinct types of models are typically used:

1. Prediction

   Goal is to identify the independent variables which best predict the outcome of interest.

   "Best" is measured by minimizing the *Prediction Error (PE)*. This measures how well the model is able to predict the outcome for a **new**, randomly selected observation that was not used in estimating the parameters of the prediction model.

   The statistical significance of individual predictors is less important.

   Screening Candidate Models: *Best subsets* variable selection procedure is preferred. It exhaustively examines models with various numbers of candidate predictors. Selected variables can be included in all models.

2. Evaluating a Predictor of Primary Interest

   a. Observational data

   The major issue in evaluating a predictor of primary interest is to rule out confounding.

   Confounding is observed when the effect of a predictor on the outcome is changed by the presence of another variable. For confounding to be observed, the confounding variable must be related to the outcome and the predictor of interest.

   A *Mediator* is a variable that is hypothesized to lie on the causal pathway between the predictor of interest and the outcome. Mediators are not included as independent variables in the model.

   Selection of potential confounders:

   1. Confounders included for face validity

      Some confounders may be well-established by previous studies. These should be included in the model, irrespective of the strength of the statistical association with the primary predictor and the outcome in the current study.

   2. Confounders identified statistically

      The recommended variable selection procedure is backwards elimination. However a more liberal criterion for removal is recommended, i.e. only removing variables with p-values $\geq$ 0.2. A comparably effective alternative is to retain variables if removing them changes the regression coefficient for the predictor of interest by more than 10% (typically used in epidemiology studies).

Interactions with the predictor of primary interest:

Typically only consider interactions between the predictor of primary interest and the confounders (not between the confounders with each other). In some situations, expect such interaction based on results from previous studies. When an interaction is identified *de novo* in a study it should be cautiously interpreted, as such analyses are susceptible to false-positive findings.

b. Randomized experiments

The intervention is the predictor of primary interest. Randomization removes confounding because it balances the confounders across the intervention groups (on average). Generally, other covariates are not included in the statistical assessment of intervention effect. There are however, two situations in which adjusting for other variables should be considered:

1. Providing valid inference in stratified designs.

2. Adjusting for baseline imbalances.

3. Identifying Multiple Important Predictors

Selecting a single best model is more difficult in this setting. Overfitting and false-positive results are more problematic, particularly for novel associations.

Still want to rule out confounding. In this context, confounding can be of concern for any of the independent predictors of interest. A preferred approach is relatively large models that include variables necessary for face validity, as well as those that meet a liberal backward selection criteria.

Still need to assess interactions. Now the number of potential interactions can be overwhelming. Assessing all potential interactions among independent variables can easily lead to false-positive findings.

All models fit in this setting should be cautiously interpreted. All associations between the independent variables and outcome require substantive interpretation. In particular, novel, implausible, weak, and borderline statistically significant associations should be cautiously interpreted.

Considerations for all models

a. Number of predictors

Model performance can be severely degraded by including too many predictors.

Rule of thumb – Ten observations for each potential predictor (not just those in the final model).

Multiple Linear Regression

Goal is to determine whether or not there is a linear relationship between a dependent variable, y, and multiple independent variable, $x_1, x_2, \ldots, x_k$

Statistical model for multiple linear regression is:

$$y = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \ldots + \beta_k \cdot x_k + e$$

Simultaneous test for whether the partial-regression coefficients are all equal to 0

H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$ H_1: At least one of the $\beta_j \neq 0$

Then the test statistic is

$$F = \frac{\text{Regression MS}}{\text{Residual MS}} = \left(\frac{\sum_{i=1}^{n} \left(\hat{y}_i - \bar{y} \right)^2 / k}{\left[\sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2 - \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y} \right)^2 \right] / (n - k - 1)} \right)$$

The p-value is p = Probability($F_{k,n\text{-}k\text{-}1} > F$)

Test for whether a particular partial-regression coefficient equals 0

H_0: $\beta_j = 0$, all other $\beta_j \neq 0$ H_1: $\beta_j \neq 0$, all other $\beta_j \neq 0$

Then the test statistic is

$$t = b_j / se(b_j)$$

The p-value is p = 2 · (area to the left of t under a t_{n-k-1} distribution) if $t < 0$
p = 2 · (area to the right of t under a t_{n-k-1} distribution) if $t \geq 0$

R^2 measures the proportion of variance in the dependent variable that is explained by the set of independent variables (also called the coefficient of determination)

Logistic Regression

Logistic regression is used determine whether or not there is a relationship between a binary dependent variable, y, and multiple independent variables, x_1, x_2, …, x_k

The logit transformation is used to create a linear function in the coefficients:

$$\text{log} - \text{odds(disease)} = \ln\left[\frac{\pi(x)}{1 - \pi(x)}\right] = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + ... + \beta_k \cdot x_k$$

Note that after the logit transform we have a model that is linear in x_1, x_2, …, x_k

The logistic regression coefficients, α, β_1, β_2, …, β_k are estimated by maximizing the likelihood (probability of the observed data as a function of the underlying parameters).

To test whether a particular regression coefficient equals 0 use the Wald statistic:

H_0: $\beta_j = 0$, all other $\beta_j \neq 0$ H_1: $\beta_j \neq 0$, all other $\beta_j \neq 0$

$$z = b_j / se(b_j) \sim N(0,1)$$

To test whether a set of coefficients are all simultaneously equal to 0 use a likelihood ratio test:

H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$ H_1: At least one of the $\beta_j \neq 0$

Likelihood ratio test = Difference in the log-likelihoods from the models with and without variable(s) being tested; a large difference implies that the null hypothesis is rejected.

Likelihood ratio = -2 ·[ln(likelihood without variable(s) - ln(likelihood with variable(s)] ~ χ^2_p
where p = number of variables removed from the model

To construct a likelihood ratio test:
1) Fit the model containing the variables we want to test and save the log-likelihood
2) Fit the model removing the variables we want to test and save the log-likelihood
3) Compare the two log-likelihoods to perform the likelihood ratio test

Cox Proportional Hazards Model

Goal is to model the time-to-event (survival) as a function of the independent variables (covariates).

Comparing survival curves using a log-rank test is a special case of a Cox proportional hazards model: single explanatory variable with two (or more) groups.

Model for the hazard rate as a function of time and the independent variables:

> Hazard rate = Probability(event occurs in a particular interval given that the event has not occurred at the beginning of the interval) = Instantaneous failure rate
>
> Let $h(t; x_1, x_2, \dots x_k)$ be the hazard rate at time t given independent variables $x_1, x_2, \dots x_k$
>
> The Cox proportional hazards model is:
>
> $$h(t; x_1, x_2, \dots x_k) = h_0(t) \cdot e^{\beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \dots + \beta_k \cdot x_k}$$
>
> where $h_0(t)$ is the baseline hazard and $\beta_1, \beta_2, \dots \beta_k$ are the parameters that assess the influence of the independent variables.

Hypothesis Tests for $\beta_1, \beta_2, \dots \beta_k$:

a) Can test whether or not a single parameter = 0 using a Wald statistic, since $\hat{\beta}$ has a normal distribution for large samples

b) Can also test whether a single parameter = 0 or a set of parameters = 0 using a likelihood ratio test

Poisson Regression Model

The Poisson distribution describes the number of events in a very large number of trials during a specified time period.

Poisson Regression Model:

Assume the count in each cell has a Poisson distribution where q_{ij} is the Poisson rate.

Poisson regression model is:

$$\log(q_{ij}) = \alpha + \beta_i + \delta_j \qquad (i = 1, 2, \ldots, m;\ j = 1, 2, \ldots, k)$$

Parameters (α, β, δ) can be estimated using maximum likelihood.

Log-Binomial Model

Goal is to determine whether or not there is a relationship between a binary dependent variable, y, and multiple independent variables, x_1, $x_2, \ldots, x_k$

An alternative to the use of logistic regression is log-binomial regression to directly estimate the relative risk in a cohort study with a more common outcome (> 10%).

$$\log(\pi) = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \ldots + \beta_k \cdot x_k$$

where the $\log(\pi)$ is the logarithm of the probability of the event.

Then the relative risk is estimated by:

$$RR = e^{\beta}$$

The log-binomial model produces unbiased estimates of the adjusted relative risk.

Major drawback is that the algorithm may not converge to provide parameter estimates.