Statistical genetics: statistical concepts in a nutshell

Content

Likelihood ratio test

- The null hypothesis $\mathcal{H}_0$ states that model $\mathcal{M}_0$ fits the observed data $\boldsymbol{y}$ as well as $\mathcal{M}_1$ does, while the alternative hypothesis $\mathcal{H}_1$ states the opposite
- The test statistic is defined as

$$t(\boldsymbol{y}) = -2 \log \left\{ \frac{p(\boldsymbol{y} \mid \hat{\boldsymbol{\theta}}_0, \mathcal{M}_0)}{p(\boldsymbol{y} \mid \hat{\boldsymbol{\theta}}_1, \mathcal{M}_1)} \right\},$$

where $\mathcal{M}_0$ and $\mathcal{M}_1$ are respectively the simpler (null) and more complex (alternative) model and $\hat{\boldsymbol{\theta}}_0$ and $\hat{\boldsymbol{\theta}}_1$ the corresponding maximum likelihood estimates

Likelihood ratio test

- The distribution of the test statistic

$$t(\boldsymbol{y}) = -2 \log \left\{ \frac{p(\boldsymbol{y} \,|\, \hat{\boldsymbol{\theta}}_0, \mathcal{M}_0)}{p(\boldsymbol{y} \,|\, \hat{\boldsymbol{\theta}}_1, \mathcal{M}_1)} \right\},$$

  under the assumption that $\mathcal{H}_0$ is true, converges against a $\chi^2(\nu)$ distribution as n approaches ∞ with ν being the difference in the number of parameters between $\mathcal{M}_1$ and $\mathcal{M}_0$
- The observed value $t(\boldsymbol{y})$ of the test statistic is calculated and the hypothesis $\mathcal{H}_0$ that model $\mathcal{M}_0$ fits the data as well as $\mathcal{M}_1$ does is rejected if $t(\boldsymbol{y}) > c$

χ^2 approximation to the likelihood ratio for a simple $\mathcal{H}_0$

- $\mathcal{H}_0 : \theta = \theta_0$ and $\mathcal{H}_1 : \theta \neq \theta_0$
- Expand the log-likelihood $\ell(\theta_0 \,|\, \boldsymbol{y}) = p(\boldsymbol{y} \,|\, \theta_0)$ as a second-order Taylor series around the maximum likelihood estimate $\hat{\theta}$

$$\ell(\theta_0 \,|\, \boldsymbol{y}) \approx \ell(\hat{\theta} \,|\, \boldsymbol{y}) + \ell'(\hat{\theta} \,|\, \boldsymbol{y})(\theta_0 - \hat{\theta}) + \ell''(\hat{\theta} \,|\, \boldsymbol{y})(\theta_0 - \hat{\theta})^2/2$$

- Plug the expansion into $t(\boldsymbol{y}) = -2\ell(\theta_0 \,|\, \boldsymbol{y}) + 2\ell(\hat{\theta} \,|\, \boldsymbol{y})$

$$\begin{aligned} t(\boldsymbol{y}) &\approx -2\ell(\hat{\theta} \,|\, \boldsymbol{y}) - \ell''(\hat{\theta} \,|\, \boldsymbol{y})(\theta_0 - \hat{\theta})^2 + 2\ell(\hat{\theta} \,|\, \boldsymbol{y}) \\ &= -\ell''(\hat{\theta} \,|\, \boldsymbol{y})(\theta_0 - \hat{\theta})^2 \end{aligned}$$

- By the LLN and since $\hat{\theta}$ is a consistent estimator

$$-\frac{1}{n}\ell''(\hat{\theta} \,|\, \boldsymbol{y}) \xrightarrow{\mathcal{P}} -\mathbb{E}[\ell''(\theta_0 \,|\, \boldsymbol{y})] = \mathcal{I}(\theta_0)$$

χ^2 approximation to the likelihood ratio for a simple $\mathcal{H}_0$ cont'd

- The maximum likelihood estimator $\hat{\theta}$ is asymptotically normal:

$$\sqrt{n\mathcal{I}(\theta_0)}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{D}} \text{Normal}(0, 1)$$

- The distribution of the test statistic $t(\boldsymbol{y})$ converges against a $\chi^2(1)$ distribution as n approaches ∞

$$n\mathcal{I}(\theta_0)(\hat{\theta} - \theta_0)^2 \xrightarrow{\mathcal{D}} \chi^2(1)$$

because the square of a standard normal random variable is χ^2 distributed with 1 degree of freedom

Bayesian information criterion (BIC)

- Approximate the prior predictive distribution under $\mathcal{M}_i$ with Laplace's method

$$p(\boldsymbol{y} \,|\, \mathcal{M}_i) \approx (2\pi)^{d/2} \det(Q)^{-1/2} p(\boldsymbol{y} \,|\, \hat{\theta}, \mathcal{M}_i) p(\hat{\theta} \,|\, \mathcal{M}_i)\,,$$

  where $\hat{\theta}$ is the maximum likelihood estimate and Q the negative Hessian of the log-likelihood evaluated at $\hat{\theta}$

- The BIC is derived by writing

$$\begin{aligned} -2 \log p(\boldsymbol{y} \,|\, \mathcal{M}_i) \approx & -d \log(2\pi) + \log \det(Q) - 2 \log p(\hat{\theta} \,|\, \mathcal{M}_i) \\ & - 2 \log p(\boldsymbol{y} \,|\, \hat{\theta}, \mathcal{M}_i) \end{aligned}$$

- By the LLN and since $\hat{\theta}$ is a consistent estimator

$$Q = -\frac{n}{n} \ell''(\hat{\theta} \,|\, \boldsymbol{y}) \xrightarrow{\mathcal{P}} -n \mathbb{E}[\ell''(\theta_0 \,|\, \boldsymbol{y})] = n \mathcal{I}(\theta_0)$$

Bayesian information criterion (BIC) cont'd

- -2 times the log prior predictive distribution is thus

$$-2\log p(\boldsymbol{y}\,|\,\mathcal{M}_i) \approx -d\log(2\pi) + d\log n + \log\det\{\mathcal{I}(\theta_0)\} - 2\log p(\hat{\theta}\,|\,\mathcal{M}_i) - 2\log p(\boldsymbol{y}\,|\,\hat{\theta},\mathcal{M}_i)$$

- Dropping all terms that remain fixed as the sample size approaches ∞ results in

$$-2\log p(\boldsymbol{y}\,|\,\mathcal{M}_i) \approx \mathrm{BIC}_i = -2\log p(\boldsymbol{y}\,|\,\hat{\theta},\mathcal{M}_i) + d\log n$$

- Small BIC values correspond to better models
- The Akaike Information Criterion is equal to

$$\mathrm{AIC}_i = -2\log p(\boldsymbol{y}\,|\,\hat{\theta},\mathcal{M}_i) + 2d$$

- Small AIC values also correspond to better models

Bayesian model averaging

- Considering that some models perform equally well, it seems reasonable to base inference on several models by using Bayesian model averaging
- Model uncertainty is then accounted for by including information from all models weighted by their posterior model probability
- The model-averaged posterior of some quantity of interest Δ with the same interpretation across models is given by

$$p(\Delta \,|\, \boldsymbol{y}) = \sum_{k=1}^{K} p(\Delta \,|\, \mathcal{M}_k, \boldsymbol{y}) p(\mathcal{M}_k \,|\, \boldsymbol{y})$$

Bayesian model averaging cont'd

- The posterior probability of model $\mathcal{M}_k$ is given by

$$p(\mathcal{M}_k \,|\, \boldsymbol{y}) = \frac{p(\boldsymbol{y} \,|\, \mathcal{M}_k) p(\mathcal{M}_k)}{\sum_{l=1}^{K} p(\boldsymbol{y} \,|\, \mathcal{M}_l) p(\mathcal{M}_l)} \,,$$

  where $p(\boldsymbol{y} \,|\, \mathcal{M}_k)$ denotes the marginal likelihood of model $\mathcal{M}_k$ and $p(\mathcal{M}_k)$ the prior probability that model $\mathcal{M}_k$ is true

- Assuming uniform prior model probabilities, the approximate posterior model probability of model $\mathcal{M}_k$ is using the BIC given by

$$\hat{p}(\mathcal{M}_k \,|\, \mathcal{D}) = \frac{\exp\left\{-\frac{1}{2}\mathsf{BIC}_k\right\}}{\sum_{l=1}^{K} \exp\left\{-\frac{1}{2}\mathsf{BIC}_l\right\}}$$

Bayesian model averaging for logistic regression

- The model-averaged posterior inclusion probability of a predictor $\boldsymbol{x}_i$ is given by

$$p(\beta_i \neq 0 \,|\, \mathcal{D}) = \sum_{k=1}^{K} \mathbb{1}_{\mathcal{M}_k}(\beta_i) p(\mathcal{M}_k \,|\, \mathcal{D})$$

The model-averaged posterior mean of β_i is given by

$$\mathbb{E}[\beta_i \,|\, \mathcal{D}] = \sum_{k=1}^{K} \mathbb{E}[\beta_i \,|\, \mathcal{M}_k, \mathcal{D}] p(\mathcal{M}_k \,|\, \mathcal{D})$$

Bayesian model averaging for logistic regression cont'd

- The model-averaged posterior variance of β_i is given by

$$\mathbb{V}[\beta_i \,|\, \mathcal{D}] = \sum_{k=1}^{K} \left\{ \left(\mathbb{V}[\beta_i \,|\, \mathcal{M}_k, \mathcal{D}] + \mathbb{E}[\beta_i \,|\, \mathcal{M}_k, \mathcal{D}]^2 \right) \times p(\mathcal{M}_k \,|\, \mathcal{D}) \right\} - \mathbb{E}[\beta_i \,|\, \mathcal{D}]^2$$

- If the sample size of the observed data $\boldsymbol{y}$ is large, then the posterior $p(\beta_i \,|\, \boldsymbol{y}, \mathcal{M}_k)$ of β_i under model $\mathcal{M}_k$ is asymptotically normal
- The mean is equal to the maximum likelihood estimator and variance equal to respective diagonal element of the inverse of the observed information matrix evaluated at the maximum likelihood estimator

Bayesian model averaging for logistic regression cont'd

- Dobutamine stress echocardiography study at the UCLA School of Medicine from 1991 until it closed in 1996
- The aim of the study was to assess if measurements taken during the stress echocardiography may be used to predict cardiac death, heart attack or coronary heart disease

Top 5 posterior model probabilities

Rank	Model	Posterior model probability†	Cumulative posterior model probability	Posterior model odds
01	$\mathcal{M}_1$: posSE, dobEF, hxofHT, restwma	0.0948	0.0948	1.00
02	$\mathcal{M}_2$: posSE, dobEF	0.0864	0.1812	1.10
03	$\mathcal{M}_3$: posSE, dobEF, restwma	0.0818	0.2631	1.16
04	$\mathcal{M}_4$: posSE, dobEF, hxofHT	0.0797	0.3427	1.19
05	$\mathcal{M}_5$: posSE, dobEF, hxofHT, ecg	0.0719	0.4147	1.32

Model comparison

Bayesian model averaging estimate

Predictor	Posterior inclusion probability	Posterior mean	Posterior standard deviation	95% equal tail interval	
				lower	upper
intercept	1.000	-0.34999	1.1031	-2.5120	1.8120
posSE	0.978	1.1126	0.2975	0.5295	1.6957
dobEF	0.882	-0.03617	0.0177	-0.0655	-0.0166
hxofHT	0.546	0.42712	0.4550	0.1586	1.4061
restwma	0.492	0.43209	0.5078	0.1647	1.5915
ecg	0.403	0.31211	0.4315	0.1419	1.4064
hxofMI	0.208	0.11074	0.2502	-0.0121	1.0750
hxofDM	0.147	0.06297	0.1811	-0.0753	0.9344
baseEF	0.132	-0.00277	0.0132	-0.0807	0.0388

Summary

- Likelihood ratio tests require nested models and rely on asymptotic approximations
- The BIC and AIC are tools that help balances model complexity and fit, which is evaluated through the maximized likelihood. The BIC prefers simpler models with small amounts of data, but becomes willing to accept more complex ones with increasing amount of data.
- Bayesian model averaging is a powerful tool to account for model uncertainty. The BIC may be used to approximate posterior model probabilities

Content

1. Motivation
2. Approaches to statistics
3. Hypothesis testing
4. Model comparison
5. **SNP association studies**

Relationship between SNP genotype and phenotype

- Genetic information is stored in the DNA in form of 4 nucleotide bases
- The human reference genome is approximately 3 giga bases long and any 2 humans differ in their genetic code by a small fraction
- Single Nucleotide Polymorphism (SNP) is a form of genetic variation at a genetic site at which the nucleotide base between 2 humans differs

Relationship between SNP genotype and phenotype cont'd

- The genotype is, among other factors, a strong influence on the phenotype
- An association between genotype and phenotype may be presumed for disease susceptibility, drug treatment or crop yields
- In case of drug treatments, some people react normally to the treatment, whereas others show none or life-threatening effects
- A particular set of SNPs may be characteristic for these phenotypes
- The goal of genome-wide association studies is then to reveal SNP patterns that permit disease susceptibility screens or personalized drug treatments

Relationship between SNP genotype and phenotype cont'd

- DNA carried by the chromosomes is present in 2 copies
- Without considering DNA copy number variation, the genotype at a biallelic SNP is either
  - AA - 2 copies of the common allele
  - AB - 1 copy of each allele
  - BB - 2 copies of the rare allele

  where allele refers to the particular nucleotide base.
- The frequency distribution of a SNP genotype and phenotype can be visualized in a contingency table

	AA	**AB**	**BB**	
Case	n_{AA}^{Case}	n_{AB}^{Case}	n_{BB}^{Case}	n^{Case}
Control	$n_{AA}^{Control}$	$n_{AB}^{Control}$	$n_{BB}^{Control}$	$n^{Control}$
	n_{AA}	n_{AB}	n_{BB}	n

General genotype count model: prospective model

- The counts may be modeled directly

	AA	**AB**	**BB**	
Case	n_{AA}^{Case}	n_{AB}^{Case}	n_{BB}^{Case}	n^{Case}
Control	$n_{AA}^{Control}$	$n_{AB}^{Control}$	$n_{BB}^{Control}$	$n^{Control}$
	n_{AA}	n_{AB}	n_{BB}	n

- In a prospective model, the phenotype is the random variable, whereas the genotype variable is supposed to be known
- Under the null hypothesis $\mathcal{H}_0$, there exists no association between both variables and thus

$$p(\boldsymbol{y} \,|\, \theta, \mathcal{H}_0) = \binom{n}{n^{Case}} \theta^{n^{Case}} (1-\theta)^{n^{Control}}$$

General genotype count model: prospective model cont'd

- Assume that the prior of θ, which represents the probability of being a case, is a Beta distribution

$$p(\theta) = \frac{1}{\mathcal{B}(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1},$$

where the Beta function is equal to

$$1/\mathcal{B}(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$$

- The prior predictive distribution is then

$$\begin{aligned} p(\boldsymbol{y} \mid \mathcal{H}_0) &= \binom{n}{n^{\text{Case}}} \frac{1}{\mathcal{B}(\alpha, \beta)} \int_0^1 \theta^{n^{\text{Case}}+\alpha-1}(1-\theta)^{n^{\text{Control}}+\beta-1} \, \mathrm{d}\theta \\ &= \binom{n}{n^{\text{Case}}} \frac{\mathcal{B}(n^{\text{Case}}+\alpha, n^{\text{Control}}+\beta)}{\mathcal{B}(\alpha, \beta)} \end{aligned}$$

General genotype count model: prospective model cont'd

- Under the alternative hypothesis $\mathcal{H}_1$, the 3 genotypes are assumed to be independent and thus

$$p(\boldsymbol{y} \,|\, \tau_{AA}, \tau_{AB}, \tau_{BB}, \mathcal{H}_1) = \binom{n}{n^{\text{Case}}} \prod_{i \in \{AA,AB,BB\}} \tau_i^{n_i^{\text{Case}}} \times (1-\tau_i)^{n_i^{\text{Control}}}$$

- Assume that the prior of τ_i, which represents the probability of being a case given that the genotype is i, has also a Beta distribution

General genotype count model: prospective model cont'd

- The prior predictive distribution is then

$$
\begin{aligned}
p(\boldsymbol{y} \mid \mathcal{H}_1) &= \binom{n}{n^{\mathsf{Case}}} \prod_{i \in \{AA,AB,BB\}} \frac{1}{\mathcal{B}(\alpha,\beta)} \times \\
&\quad \int_0^1 \tau_i^{n_i^{\mathsf{Case}}+\alpha-1} (1-\tau_i)^{n_i^{\mathsf{Control}}+\beta-1} \, \mathrm{d}\tau_i \\
&= \binom{n}{n^{\mathsf{Case}}} \prod_{i \in \{AA,AB,BB\}} \frac{\mathcal{B}(n_i^{\mathsf{Case}}+\alpha, n_i^{\mathsf{Control}}+\beta)}{\mathcal{B}(\alpha,\beta)}
\end{aligned}
$$

General genotype count model: prospective model cont'd

- The prospective Bayes factor is

$$
\begin{aligned}
B_{10} &= \frac{p(\boldsymbol{y} \,|\, \mathcal{H}_1)}{p(\boldsymbol{y} \,|\, \mathcal{H}_0)} \\
&= \frac{\mathcal{B}(\alpha, \beta)}{\mathcal{B}(n^{\text{Case}} + \alpha, n^{\text{Control}} + \beta)} \times \\
&\quad \prod_{i \in \{AA, AB, BB\}} \frac{\mathcal{B}(n_i^{\text{Case}} + \alpha, n_i^{\text{Control}} + \beta)}{\mathcal{B}(\alpha, \beta)}
\end{aligned}
$$

- Data-dependent hyperparameters may be used:

$$
(\alpha, \beta) = \lambda \left(n^{\text{Case}}/n, n^{\text{Control}}/n \right) ,
$$

  which are uninformative in distinguishing $\mathcal{H}_0$ from $\mathcal{H}_1$ and where λ is used to scale the effect size.

Thank you for your attention