STATS 305B Notes

Kenneth Tay

Contents

1 Estimating/Testing a Binomial Parameter

Say $Y \sim \text{Bin}(n, \pi)$, we wish to estimate π.

- MLE is $\hat{\pi} = y/n$.
- $\mathbb{E}\hat{\pi} = \pi$, Var $\hat{\pi} = n\pi(1-\pi)$.
- **Wald test:** $\hat{\pi} \approx \mathcal{N}\left(\pi_0, (I(\hat{\pi})^{-1})\right)$, so the corresponding z-statistic is $z_{Wald} = \dfrac{\hat{\pi} - \pi_0}{\sqrt{\hat{\pi}(1-\hat{\pi})/n}}$. The $(1-\alpha)$-level confidence interval is given by

$$\left(\hat{\pi} - z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}, \hat{\pi} + z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \right).$$

- **Score test:** Let $S = \left.\dfrac{\partial L(\pi \mid Y)}{\partial \pi}\right|_{\pi_0}$. Under H_0, $S \approx \mathcal{N}(0, I(\pi_0))$, and the corresponding z-statistic is

$z_{score} = \dfrac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}$. The $(1-\alpha)$-confidence interval is given by

$$\left\{ \pi : \frac{(\hat{\pi} - \pi)^2}{\pi(1-\pi)/n} \leq \chi^2_{1,1-\alpha} \right\}.$$

- **Likelihood ratio test:** The binomial log-likelihood function is equal to $L_0 = y \log \pi_0 + (n-y)\log(1-\pi_0)$ under H_0, and is equal to $L_1 = y \log \hat{\pi} + (n-y)\log(1-\hat{\pi})$ more generally. Hence, the likelihood-ratio test statistic simplifies:

$$-2(L_0 - L_1) = 2\left[y \log \frac{\hat{\pi}}{\pi_0} + (n-y)\log\frac{1-\hat{\pi}}{1-\pi_0} \right],$$

  which has approximate distribution χ^2_1 under H_0. The $(1-\alpha)$-level confidence interval is given by

$$\left\{ \pi : -2(L_0 - L_1) \leq \chi^2_{1,1-\alpha} \right\}.$$

2 Estimating/Testing Multinomial Parameters

Suppose $Y \sim \text{Multinom}(n, \pi)$, where $\pi \in \mathbb{R}^k$ represents the probability of Y being in each of k categories.

- If we let $\Sigma(\pi) = \text{diag}(\pi) - \pi\pi^T$, then we have $\text{Cov}(Y) = n\Sigma(\pi)$. (Note: The rank of $\Sigma(\pi)$ is $k-1$.) As $n \to \infty$, $\dfrac{Y}{n} - \pi = \hat{\pi}_{MLE} - \pi \sim \mathcal{N}\left(0, \dfrac{\Sigma(\pi)}{n}\right)$.

- **Pearson χ^2 test:** To test $H_0 : \pi = \pi_0$, the test statistic is $\displaystyle\sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j} = \sum_{j=1}^{k} \frac{(y_j - n\pi_{0,j})^2}{n\pi_{0,j}}$, which has an approximate χ^2_{k-1} distribution under H_0.

- **Likelihood ratio test:** To test $H_0 : \pi = \pi_0$, the test statistic is $G(\pi_0, \hat{\pi}) = 2\displaystyle\sum_{j=1}^{k} y_j \log\left(\frac{y_j/n}{\pi_{0,j}}\right)$, which has an approximate χ^2_{k-1} distribution under H_0.

3 Contingency Tables

A contingency table is a tabulation of the empirical joint distribution of categorical random variables.

Let (X, Y) be categorical random variables with I and J categories respectively. Define

- $Y_{ij} :=$ number of observations in cell (i, j).
- $\pi_{ij} := P\{X = i, Y = j\}$ (joint probabilities).
- $\pi_{i+} := P\{X = i\}$, $\pi_{+j} := P\{Y = j\}$ (marginal probabilities).
- $\pi_{j|i} = P\{Y = j \mid X = i\} = \pi_{ij}/\pi_{i+}$ (conditional probabilities).

3.1 Sampling models

- **Poisson sampling model.** Fix a λ, and let $N \sim \text{Poisson}(\lambda)$. Then, get a simple random sample of size N and check which cell of the contingency table each subject falls in.

  It can be shown that $Y_{ij} \overset{ind.}{\sim} \text{Poisson}(\lambda_{ij})$.
- **Multinomial sampling model.** Instead of letting the sample size be Poisson-distributed, let it be some fixed number n. (Row and column counts are not fixed.) Then we have $Y \sim \text{Multinom}\left(n, (\pi_{ij})\right)$.
- **Independent multinomial sampling model.** This can be used when either row or column totals are fixed. If row totals are fixed, then we can model the counts by $Y[i,] \sim \text{Multinom}(Y_{i+}, \pi_{\cdot|i})$. (Similar set-up if column totals are fixed.
- **Hypergeometric sampling model.** This is used when both row and column totals are fixed. See Agresti Sec 3.5.1 for details.

3.2 Comparing 2 proportions in 2×2 tables

Define $\pi_1 := \theta_{Yes|1}$, $\pi_2 := \theta_{Yes|2}$. Then there are few different ways to measure the relationship between π_1 and π_2:

- **Difference of proportions** $\pi_1 - \pi_2$.
  - **Normal approximation:** $\hat{\pi}_1 - \hat{\pi}_2 \approx \mathcal{N}\left(\pi_1 - \pi_2, \dfrac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \dfrac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}\right)$.
- **Relative risk** $r := \dfrac{\pi_1}{\pi_2}$.
  - **Normal approximation:** Usually look at $\log r$ instead. $\log\left(\dfrac{\hat{\pi}_1}{\hat{\pi}_2}\right) \approx \mathcal{N}\left(\log\left(\dfrac{\pi_1}{\pi_2}\right), \dfrac{1-\hat{\pi}_1}{y_1} + \dfrac{1-\hat{\pi}_2}{y_2}\right)$.
- **Odds ratio** $\theta := \dfrac{Odds(\pi_1)}{Odds(\pi_2)} := \dfrac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \dfrac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$.
  - In the Poisson and Multinomial sampling models, $\theta = 1$ is equivalent to X and Y being independent.
  - In the independent multinomial sampling model, $\theta = 1$ and $r = 1$ individually imply that $\pi_1 = \pi_2$.
  - Odds ratio does not change value when rows and columns are switched. This is useful for case control studies (Agresti Sec 2.2.6).
  - **Rare disease hypothesis:** When π_1 and π_2 are small, $\theta \approx r$.
  - For $2 \times 2 \times K$ tables, we can talk about **conditional odds ratios** for each level of Z. These are denoted by $\theta_{XY(1)}, \ldots, \theta_{XY(K)}$.
    - **Homogeneous** XY **association** means $\theta_{XY(1)} = \cdots = \theta_{XY(K)}$. If there is homogeneous association, we say that there is no interaction between X and Y.
    - Conditional independence is the special case of the above, where all the conditional odds ratios are 1.
    - **Collapsibility:** When there is homogeneous association, then $\theta_{XY} = \theta_{XY(K)}$ if either Z and X conditionally independent or if Z and Y conditionally independent.

- Due to convergence issues, we look at $\log\hat{\theta}$, then exponentiate to get results for $\hat{\theta}$. For large samples, we have

  $$\log\hat{\theta} \approx \mathcal{N}\left(\log\theta, \frac{1}{Y_{11}} + \frac{1}{Y_{12}} + \frac{1}{Y_{21}} + \frac{1}{Y_{22}}\right).$$

  We can get Wald confidence intervals for $\log\theta$ and θ from the above.

3.3 Testing

See Agresti Sec 6.6.4 for power calculations for χ^2 tests in contingency tables.

Tests of independence

- Purpose: To test independence of X and Y, i.e. $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$. Assume sampling model is $n \sim \text{Pois}(\mu)$. Then the entries of the table are independent with $n_{ij} \overset{ind}{\sim} \text{Pois}(\mu\pi_{ij})$.
- **Pearson χ^2 test statistic** is $X^2 = \sum_{i,j} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$, where $\hat{\mu}_{ij} = n\hat{\pi}_{i+}\hat{\pi}_{+j}$ is the MLE under the null. (This turns out to be the score statistic.)
- **Pearson residual** $e_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}}}$, **standardized residual** $r_{ij} = \frac{e_{ij}}{\sqrt{(1 - p_{i+})(1 - p_{+j})}}$. r_{ij} has asymptotic $\mathcal{N}(0, 1)$ distribution.
- **Likelihood ratio test statistic** is $G^2 = -2\log\Lambda = 2\sum_{i,j} n_{ij}\log(n_{ij}/\hat{\mu}_{ij})$.
- Both X^2 and G^2 have asymptotic χ^2 distributions with $df = (I - 1)(J - 1)$. Also, $X^2 - G^2 \xrightarrow{P} 0$.
- See Agresti Sec 3.2.3 for discussion on the adequacy of the χ^2 distribution, and Sec 3.3.7 on limitations of χ^2 tests.
- If X and Y can be treated ordinally, we may use a linear trend alternative to independence (see Agresti Sec 3.4.1 for details).

McNemar's test for marginal homogeneity in 2×2 tables (Agresti p227)

- Purpose: To test for marginal homogeneity in 2×2 tables, i.e. $H_0 : \pi_{1+} = \pi_{+1}$, or equivalently, $H_0 : \pi_{12} = \pi_{21}$.
- Under H_0, the score test statistic is $z_0 = \frac{n_{21} - n_{12}}{\sqrt{n_{21} + n_{12}}}$.

  z_0 has an asymptotic $\mathcal{N}(0, 1)$ distribution, and z_0^2 has chi-squared distribution with $df = 1$. (We typically look at z_0^2.)
- McNemar's test can be viewed as a CMH test for a $2 \times 2 \times n$ table, where each table corresponds to one paired observation.

Cochran-Mantel-Haenszel (CMH) Test of Conditional Independence (Agresti p227)

- Purpose: A non-model-based test of H_0 : conditional independence in $2 \times 2 \times K$ tables, i.e. X and Y are independent given Z.
- This is frequently used in matched pairs. Each pair is a partial table (so there are K pairs), one subject is assigned the control while the other is assigned the new treatment. The null hypothesis is that "accounting for the pair", treatment and response are independent (i.e. new treatment has no effect).
- Let the partial counts in table k be $n_{11k}, \ldots, n_{22k}$. Under the null, the hypergeometric mean and variance of n_{11k} are

$$\begin{aligned} \mu_{11k} = \mathbb{E}n_{11k} &= \frac{n_{1+k}n_{+1k}}{n_{++k}}, \\ \text{Var } n_{11k} &= \frac{n_{1+k}n_{2+k}n_{+1k}n_{+2k}}{n_{++k}^2(n_{++k}-1)}. \end{aligned}$$

- The test statistic is

$$CMH = \frac{\left[\sum_k (n_{11k} - \mu_{11k})\right]^2}{\sum_k \text{Var } n_{11k}}.$$

It has large-sample χ^2 distribution with $df = 1$.

- When $K = 1$, this is the same as the Pearson χ^2 statistic for the 2×2 table.

Fisher's exact test for 2×2 tables (Agresti p90)

- Purpose: For testing H_0 : independence, in the case of fixing both row and column totals.
- In this case, $\mathbb{P}(n_{11} = t) = \dfrac{\binom{n_{1+}}{t}\binom{n_{2+}}{n_{+1}-t}}{\binom{n}{n_{+1}}}$ has a hypergeometric distribution.

3.4 Linear logit model for $I \times 2$ tables (Agresti Sec 5.3.4)

- We can use this if the categories for X have some natural ordering.
- Model: $\text{logit}(\pi_i) = \alpha + \beta x_i$, where $(x_1, \ldots, x_I)$ are the scores for the categories of X. (Natural extension to multiway contingency tables in Agresti Sec 5.4.1.)
- Independence corresponds to $\beta = 0$. To test for this (in the case where each row can be thought of as an independent $\text{Binom}(n_i, \pi_i)$ realization), we can use the **Cochran-Armitage Trend Test** (Agresti p178). The test statistic is given by

$$z^2 = \left[\frac{\sum_i (x_i - \bar{x})y_i}{\sqrt{p(1-p)\sum_i n_i(x_i - \bar{x})^2}}\right]^2,$$

where $p = \left(\sum_i y_i\right)/n$ (i.e. overall proportion of successes). Under the null, it is asymptotically χ^2 with $df = 1$.

4 Generalized Linear Models (GLMs)

4.1 Set-up

A GLM consists of 3 parts:

1. **The random component:** Responses $Y_1, \ldots, Y_n$ are independent observations from a distribution in the exponential family, taking the form

$$f_Y(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\},$$

   for specific functions a, b and c. This is a 1- or 2-parameter exponential family, depending of whether ϕ is known or unknown.

   - Let $\ell(\theta,\phi;y) = \log f_Y(y;\theta,\phi)$ be the log-likelihood function.
   - (M&N p29, Agresti p131) $\mathbb{E}Y = \mu = b'(\theta)$, and $\text{Var } Y = b''(\theta)a(\phi)$. We often write $b''(\theta) = V(\mu)$ and call it the **variance function**.
   - $a(\phi)$ commonly of the form $a(\phi) = \phi/w$. Here, ϕ is called the **dispersion parameter** and is sometimes denoted by σ^2.

2. **The systematic component:** Covariates $x_1, \ldots, x_p$ produce a linear predictor η given by $\eta = \sum_{j=1}^{p} x_j\beta_j$.

3. **The link function:** A function g such that $\eta_i = g(\mu_i)$, where $\mu_i = \mathbb{E}Y_i$.

   - Logit link: $\eta = \log\left(\frac{\mu}{1-\mu}\right)$.
   - Probit link: $\eta = \Phi^{-1}(\mu)$.
   - Complementary log-log link: $\eta = \log[-\log(1-\mu)]$.
   - Power family of links: $\eta = (\mu^\lambda - 1)/\lambda$ if $\lambda \neq 0$, $\eta = \log\mu$ if $\lambda = 0$.
   - **Canonical link:** The link function when $\eta = \theta$ (i.e. the canonical parameter in the random component). For canonical links, the sufficient statistic is X^TY (in vector notation).

4.2 Assumptions

- Given the predictors X_i, the Y_i are independent.
- Y_i's follow the stated parametric form (in the random component).
- The link function has the stated form.

4.3 Fitting the model

Maximum likelihood

- Differentiating the log-likelihood and setting it to zero, we obtain the **score equation** $X^Ty - X^T\hat{\mu} = 0$. In general, $\hat{\mu}$ is a non-linear function of $\hat{\beta}$, so the score equation cannot be solved directly. (For details, see Agresti Sec 4.4.5 p133.)

- (Agresti Sec 4.4.8, p135) We can determine the asymptotic covariance matrix of the MLE $\hat{\beta}$. Let W be a diagonal matrix with diagonal elements $w_i = \dfrac{1}{\text{Var } Y_i}\left(\dfrac{\partial \mu_i}{\partial \eta_i}\right)^2$. Then the inverse of the asymptotic covariance matrix, also known as the **information matrix**, is $\mathcal{J} = X^T W X$.
- Note that W depends on the link function. $\mathcal{J}$ will be used in fitting procedures.

Newton-Raphson method (Agresti Sec 4.6.1)

- Given a first guess, it obtains a second guess by approximating the function in a neighborhood of the first guess by a second-degree polynomial, and then finding the location of that polynomial's maximum.
- Let $L(\beta)$ be the function to be maximized. Let $u = \nabla L = \left(\frac{\partial L(\beta)}{\partial \beta_0}, \frac{\partial L(\beta)}{\partial \beta_1}, \dots\right)^T$. Let H be the Hessian.

  If our current guess is $\beta^{(t)}$, then the next guess is $\beta^{(t+1)} = \beta^{(t)} - \left[H^{(t)}\right]^{-1} u^{(t)}$, where $H^{(t)}$ and $u^{(t)}$ are the Hessian and score evaluated at the current guess.
- Newton-Raphson uses **observed information**, i.e. $H_{ij} = \dfrac{\partial^2 L(\beta)}{\partial \beta_i \partial \beta_j}$ evaulated at $\beta^{(t)}$.

Fisher scoring (Agresti Sec 4.6.2)

- Instead of observed information, Fisher scoring uses **expected information**.
- Let $\mathcal{J}$ have elements $\mathcal{J}_{ij} = -\mathbb{E}\left[\dfrac{\partial^2 L(\beta)}{\partial \beta_i \partial \beta_j}\right]$. Let $\mathcal{J}^{(t)}$ be $\mathcal{J}$ evaluated at the current guess $\beta(t)$. The the next guess is $\beta^{(t+1)} = \beta^{(t)} + \left[\mathcal{J}^{(t)}\right]^{-1} u^{(t)}$.
- For GLMs with canonical links, observed and expected information are the same.

Iterated reweighted least squares (IRLS)

- Given a current guess $\hat{\beta}^{(t)}$ for the parameters, consider the linear model with the correct mean and correct variance: $y_i = \mu_i(\beta) + \varepsilon_i$, with $\varepsilon_i \sim \mathcal{N}(0, W(\hat{\eta}_i^{(t)}))$.
- Using a Taylor series approximation for $\mu(\beta)$, we get the approximate weighted linear model $y = \hat{\mu}^{(t)} + \hat{W}^{(t)}(X\beta - X\hat{\beta}^{(t)}) + \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, \hat{W}^{(t)})$, where $\hat{W}^{(t)} = \text{diag}(W(\hat{\eta}_i^{(t)}))$.
- Rearranging, we get $\hat{z}^{(t)} = X\beta + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, (\hat{W}^{(t)})^{-1})$, where $\hat{z}^{(t)} = X\hat{\beta}^{(t)} + (\hat{W}^{(t)})^{-1}(y - \hat{\mu}^{(t)})$.
- We can fit this model with usual OLS! The solution is $\hat{\beta}^{(t+1)} = (X^T \hat{W}^{(t)} X)^{-1} X^T \hat{W}^{(t)} \hat{z}^{(t)}$.
- IRLS is equivalent to Fisher scoring.

4.4 Goodness of fit / inference

- **Generalized hat matrix:** $H = W^{1/2} X (X^T W X)^{-1} X^T W^{1/2}$.
- **Asymptotic distribution** for $\hat{\beta}$ is $\hat{\beta} \sim \mathcal{N}(\beta, (X^T W X)^{-1})$. Can use this for standard errors and confidence intervals. (Note that W depends on β, so we can just use the plug-in estimate where necessary.)

- For a GLM with observations $y = (y_1, \dots, y_n)$, let $\ell(\mu; y)$ denote the log-likelihood function expressed in terms of the means $\mu = (\mu_1, \dots, \mu_n)$. Let $\ell(\hat{\mu}; y)$ denote the maximum of the log-likelihood for the GLM, and let $\ell(y; y)$ denote the maximum achievable log-likelihood, i.e. the saturated model, where the number of parameters is equal to the number of observations and has the perfect fit $\hat{\mu} = y$.
- **Scaled deviance** $D^*(y; \hat{\mu}) := 2\ell(y; y) - 2\ell(\hat{\mu}; y)$. Asymptotically, $D^*(y; \hat{\mu}) \overset{\cdot}{\sim} \chi^2_{n-p}$.
- More generally, to test model M_1 vs. model M_2, we would use the test statistic $D^*(y; \mu(\hat{\beta}_{M_1})) - D^*(y; \mu(\hat{\beta}_{M_2})) \approx \chi^2_{|M_2|-|M_1|}$.
- If we let $\hat{\theta}$ and $\tilde{\theta}$ be the estimates of the canonical parameters under the GLM and the saturated model respectively, if $a_i(\phi) = \phi/w_i$, the scaled deviance can be written as

$$D^*(y; \hat{\mu}) = \frac{2w_i[y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)]}{\phi} =: \frac{D(y; \hat{\mu})}{\phi}.$$

  $D(y; \hat{\mu})$ is called the **deviance**.
- **Generalized Pearson X^2 statistic:** $X^2 = \displaystyle\sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$.
- **McFadden's R^2:** An analog for R^2 in OLS. It is equal to $\dfrac{DEV(M_0) - DEV(M)}{DEV(M_0)}$, where M_0 is the null model (i.e. intercept-only).
- For binary response, we could set up the confusion matrix and use that to estimate predictive power.
- **Pearson residuals:** $e_i = \dfrac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$. We have $\sum e_i^2 = X^2$.
- **Deviance residuals:** Let $d_i = 2w_i[y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)]$. (Note that $\sum d_i = D(y; \hat{\mu})$.) The deviance residual is $\sqrt{d_i} \cdot \text{sign}(y_i - \hat{\mu}_i)$.
- **Standardized residuals:** Divide the Pearson residuals by their standard error: $r_i = \dfrac{e_i}{\sqrt{1 - \hat{h}_i}}$, where $\hat{h}_i$ is the estimated diagonal element of $H = W^{1/2}X(X^TWX)^{-1}X^TW^{1/2}$, the generalized hat matrix. $\hat{h}_i$ is also called the **leverage** of observation i. (Derivation in Agresti p142.)
- Can plot residuals against each predictor X_j. If there seems to be some structure, then maybe higher-order terms or interaction terms should be added.

4.5 Latent variable interpretation for binary response (Agresti Sec 4.2.6, p122)

- A subject has response $Y = 0$ or 1.
- Let $X = x$ be the value of the predictor value. Suppose that there is a threshold T such that $Y = 1 \Leftrightarrow T \leq x$.
- The threshold T varies from subject to subject, having distribution with CDF F.
- Thus, for a fixed dosage x, the probability a randomly selected subject dies is $\pi(x) = \mathbb{P}(Y = 1 \mid X = x) = \mathbb{P}(T \leq x) = F(x)$.
- In practice, we don't observe T. We assume that F belongs to some parametric family. If Φ is the standard CDF for the family, then we can write $\Phi^{-1}[\pi(x)] = \alpha + \beta x$.

4.6 Model selection

- **Forward stepwise:** Add terms sequentially. At each stage, select the term with the greatest improvement in fit, stop when adding the next predictor doesn't help fit. **Backward stepwise** does the opposite.
- **Akaike information criterion (AIC):** -2 (maximized log likelihood - number of parameters in model). The smaller the AIC, the better.
- **Bayesian information criterion (BIC):** The coefficient in front of number of parameters changes from 2 to $\log n$.

5 Logistic Regression (Agresti Ch 5)

5.1 Set-up

- **Model:** $\text{logit}(\pi) = x^T\beta$, i.e. $\log\left(\frac{\pi}{1-\pi}\right) = x^T\beta$, where $\pi = \pi_\beta(X) = P\{Y = 1 \mid X_1 = x_1, \ldots, X_p = x_p\} = \mathbb{E}_\beta[Y \mid X]$.
- From the above, we get

$$\mathbb{E}_\beta[Y \mid X] = \pi = \frac{\exp(x^T\beta)}{1+\exp(x^T\beta)}, \qquad 1-\pi = \frac{1}{1+\exp(x^T\beta)}.$$

- Since each Y_i is a Bernoulli random variable, the variance of π is given by

$$\begin{aligned}\text{Var}_\beta[Y \mid X] &= \text{diag}(\mathbb{E}_\beta[Y \mid X](1-\mathbb{E}_\beta[Y \mid X])) = \text{diag}(\pi_\beta(X)(1-\pi_\beta(X))) \\ &= \text{diag}\frac{\exp(x^T\beta)}{[1+\exp(x^T\beta)]^2} =: W_\beta(X).\end{aligned}$$

5.2 Fitting the model

- **Log-likelihood** $L = \sum_{i=1}^n Y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + \log(1-\pi_i) = (X\beta)^T Y - \sum_{i=1}^n \log\left(1 + e^{X_i^T\beta}\right).$
- **Score function:** For each $j = 1, \ldots, p$,

$$\frac{\partial \log L(\pi \mid Y)}{\partial \beta_j} = \sum_{i=1}^n X_{ij}Y_i - \frac{X_{ij}\exp(X_i^T\beta)}{1+\exp(X_i^T\beta)}.$$

Combining and writing it as a column vector:

$$\begin{aligned}\nabla \log L(\pi \mid Y) &= X^TY - X^T\frac{\exp(X^T\beta)}{1+\exp(X^T\beta)} \\ &= X^T\left(Y - \mathbb{E}_\beta[Y \mid X]\right).\end{aligned}$$

- **Hessian:** $\nabla^2 \log L(\pi \mid Y) = -X^T\text{Var}_\beta[Y \mid X]X.$
- Details on Newton-Raphson in Agresti p194-195.

- Problems with fitting may result if there is **complete/quasi-complete separation** in the data. In this case, some of the MLEs may be infinite. (When there is no separation, MLEs are finite and unique.)
  - When this happens, we usually see huge reported standard errors.
  - With infinite estimates, Wald inference cannot be done, but likelihood-ratio and score tests (and CIs) still work.
  - Fixes: Use a Bayesian approach, combine categories, regularize, drop predictors (if you suspect overfitting). See Agresti p236-237 for details.

5.3 Inference and testing

- See Agresti p138-139 for deviance of grouped vs. ungrouped data.

5.4 Goodness of fit

- **X^2 or G^2 test:** If number of levels of x is fixed, we can use the X^2 and G^2 statistics as in contingency tables.
  - The X^2 and G^2 statistics have limiting χ^2 distributions; the degrees of freedom is the number of parameters in the saturated model minus the number of parameters in the model.
  - When the x values are not grouped or are continuous, we can group them, and then apply this test.
- (Agresti p173) **Hosmer-Lemeshow test:** Partition the data into groups based on the fitted values.
  - Calculate the fitted values $\hat{\pi}$, order them, then group them into g groups. Let y_{ij} be the binary outcome for observation j in group i, and let $\hat{\pi}_{ij}$ be the corresponding fitted probability. The test statistic is
$$\sum_{i=1}^{g} \frac{\left(\sum_j y_{ij} - \sum_j \hat{\pi}_{ij}\right)^2}{\left(\sum_j \hat{\pi}_{ij}\right)\left[1 - \left(\sum_j \hat{\pi}_{ij}\right)/n\right]}.$$
  - Test statistic does not have a limiting χ^2 distribution, but when the number of distinct patterns of covariate values equals the sample size, the null distribution is approximately χ^2 with $df = g - 2$.

6 Multinomial Logistic Regression (Agresti Ch 8)

Logistic regression models capture conditional distributions $Y \mid X, Z$, so are appropriate for cases where Y is viewed as a response while X and Z and viewed as explanatory.

Let Y be a response variable with J categories. Let $\pi_j(x) = \mathbb{P}(Y = j \mid x)$.

6.1 Nominal responses

- **Baseline-category logit model:** $\log \dfrac{\pi_j(x)}{\pi_J(x)} = \alpha_j + \beta_j^T x$, $j = 1, \dots, J-1$.
- See Agresti Sec 8.1.4 for details on fitting.

6.2 Ordinal responses

- **Cumulative logit model:** $\text{logit}[\mathbb{P}(Y \leq j \mid x)] = \alpha_j + \beta^T x,\ j = 1, \ldots, J-1$. Also called the **proportional odds model**.

- **Cumulative link model:** $G^{-1}[\mathbb{P}(Y \leq j \mid x)] = \alpha_j + \beta^T x$, where G^{-1} is a link function (i.e. inverse of continuous CDF G).

- **Adjacent-categories logit model:** $\log \dfrac{\pi_j(x)}{\pi_{j+1}(x)} = \alpha_j + \beta^T x$, $j = 1, \ldots, J-1$.

6.3 Testing conditional independence in $I \times J \times K$ tables using multinomial models

See Agresti Sec 8.4 (p314-316) for details.

7 Loglinear Models for Contingency Tables (Agresti Ch 9)

Loglinear models capture the joint distribution of (X, Y, Z), so are appropriate when we want to view all variables as response variables.

- **Saturated model:** $\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$, with identifiability constraints. There are IJ parameters in this model.

  For 3 variables, the saturated model is $\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}$. Degrees of freedom is 0 (IJK parameters to be fit). The symbol for this model is (XYZ).

- **Mutual independence model:** $\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$, with identifiability constraints such as $\lambda_I^X = \lambda_J^Y = \lambda_K^Z = 0$ or $\sum_i \lambda_i^X = \sum_j \lambda_j^Y = \sum_j \lambda_k^Z = 0$. Degrees of freedom is $IJK - [(I-1) + (J-1) + (K-1) + 1] = IJK - I - J - K + 2$. The symbol for this model is (X, Y, Z).

- **Joint independence model:** Y is jointly independent of X and Z if $\pi_{ijk} = \pi_{i+k}\pi_{+j+}$ for all i, j, k. The model for this is $\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ}$. Degrees of freedom is $(J-1)(IK-1)$. The symbol for this model is (XZ, Y).

- **Conditional independence model:** If X and Y are conditionally independent given Z, the model for this is $\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$. Degrees of freedom is $(I-1)(J-1)K$. The symbol for this model is (XZ, YZ).

- **Homogeneous association model:** $\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$. Degrees of freedom is $(IJK-1) - [(I-1) + (J-1) + (K-1) + (I-1)(J-1) + (J-1)(K-1) + (I-1)(K-1)] = (I-1)(J-1)(K-1)$. The symbol for this model is (XY, XZ, YZ).

Inference

- X^2 or G^2 test could be used for goodness-of-fit.

- See Agresti Table 9.11 (p355) for equivalence between loglinear and logistic models.

- See Agresti Table 9.12 (p357) for the minimal sufficient statistics for each loglinear model.

- See Agresti Sec 9.6.2 (p357-358) and Table 9.13 (p359) for the likelihood equations for fitting the loglinear models.
- For testing **marginal homogeneity**, compare the fit under the symmetry model and the quasi-symmetry model. See Agresti Sec 11.4.3 for details.
- **Residuals:** Look at the Pearson residual $e_i = \dfrac{n_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$ (n_i observed in cell i, $\hat{\mu}_i$ fitted value for cell i), or the standardized residual $r_i = \dfrac{e_i}{\sqrt{1 - \hat{h}_i}}$, where $\hat{h}_i$ is a diagonal element of the estimated hat matrix. Standardized residuals have asymptotic $\mathcal{N}(0, 1)$ distribution.

Fitting

- Loglinear models can be fit using Newton-Raphson or iterative proportional fitting (Agresti Sec 9.7).

8 Models for Matched Pairs (Agresti Ch 11)

Here, the row and column categories are the same. For a matched pair randomly selected from the population of interest, let π_{ab} denote the probability of outcome a for the first observation and outcome b for the second observation. We then treat $\{n_{ab}\}$ as a sample from Multinom$(n; \{\pi_{ab}\})$.

Comparing proportions for 2×2 square tables

- See Agresti p414 for a confidence interval for $\pi_{+1} - \pi_{1+}$.
- **Marginal homoegeneity** refers to $\pi_{1+} = \pi_{+1}$ (which in turn implies $\pi_{2+} = \pi_{+2}$). This can be tested using **McNemar's test** (see Agresti p415).
- **McNemar-CMH test connection:** An alternative representation of binary responses for n matched pairs is n 2×2 tables, 1 for each pair. Using the CMH statistic to test for conditional independence is algebraically the chi-squared form of McNemar's statistic.

Logistic Models for 2×2 square tables

Let (Y_{i1}, Y_{i2}) denote the pair of observations for subject i, with Y_{ij} being 0/1-valued. Let $P(Y_t = 1)$ be the mean of $P(Y_{it} = 1)$. Let $x_1 = 0$, $x_2 = 1$.

- **Marginal model for matched pairs:** $P(Y_t = 1) = \alpha + \delta x_t$, or logit$[P(Y_t = 1)] = \alpha + \beta x_t$.
- **Subject-specific model:** logit$[P(Y_{it} = 1)] = \alpha_i + \beta x_t$.

Models for square tables

Let $x_1 = 0$, $x_2 = 1$.

- **Baseline-category logit model:**

$$\log\left[\frac{P(Y_t = j)}{P(Y_t = I)}\right] = \alpha_j + \beta_j x_t, \quad t = 1, 2, \quad j = 1, \ldots, I-1.$$

  - This can be used for nominal-scale matched-pair responses.
  - This model has $2(I-1)$ parameters.
  - Marginal homogeneity is the special case $\beta_1 = \cdots = \beta_{I-1} = 0$.
  - Goodness-of-fit can be tested with X^2 or G^2, with $df = I - 1$.

- **Cumulative logit model:**

$$\text{logit}\left[P(Y_t \leq j)\right] = \alpha_j + \beta x_t, \quad t = 1, 2, \quad j = 1, \ldots, I-1.$$

  - This can be used for ordinal-scale matched-pair responses.
  - This model has I parameters.
  - Marginal homogeneity is the special case $\beta = 0$.
  - Goodness-of-fit can be tested with X^2 or G^2, with $df = I - 2$.

- **Symmetry:** $I \times I$ joint distribution $\{\pi_{ab}\}$ has symmetry if $\pi_{ab} = \pi_{ba}$ for all $a \neq b$.
  - Model: $\log \mu_{ab} = \lambda + \lambda_a + \lambda_b + \lambda_{ab}$, with $\lambda_{ab} = \lambda_{ba}$.
  - Solution: $\hat{\mu}_{ab} = \dfrac{n_{ab} + n_{ba}}{2}$ for all a, b.
  - Goodness of fit test: $df = I(I-1)/2$.

- **Quasi-symmetry:**
  - Model: $\log \mu_{ab} = \lambda + \lambda_a^X + \lambda_b^Y + \lambda_{ab}$, with $\lambda_{ab} = \lambda_{ba}$.
  - Symmetry is the special case of $\lambda_a^X = \lambda_a^Y$ for $a = 1, \ldots, I$.
  - Independence is the special case of $\lambda_{ab} = 0$ for all a, b.
  - Identifiability requires further constraints, such as $\lambda_I^X = 0$ and all $\lambda_b^Y = 0$.
  - Goodness of fit test: $df = (I-1)(I-2)/2$.

- **Quasi-independence:** Variables are independent, given that the row and column outcomes differ.
  - Model: $\log \mu_{ab} = \lambda + \lambda_a^X + \lambda_b^Y + \delta_a I(a = b)$.
  - This is a special case of quasi-symmetry (see Agresti p430).
  - Goodness of fit test: $df = (I-1)^2 - I$ (for $I \geq 3$).

- **Bradley-Terry model:** Let Π_{ab} denote the probability that a is preferred to b. Assume that ties cannot occur, so $\Pi_{ab} + \Pi_{ba} = 1$.
  - Model: $\log(\Pi_{ab}/\Pi_{ba}) = \beta_a - \beta_b$.
  - We can think of β_a as some rating for how good team a is.
  - The Bradley-Terry model is a logistic formulation of the quasi-symmetry model, with $\beta_a = \lambda_a^X - \lambda_a^Y$.

Measuring agreement between observers

- Perfect agreement: $\sum \pi_{aa} = 1$.
- Can try to fit one of the models above.
- **Cohen's kappa:** Compares the probability of agreement $\sum \pi_{aa}$ to that expected if the ratings were independent:
$$\kappa = \frac{\sum \pi_{aa} - \sum \pi_{a+}\pi_{+a}}{1 - \sum \pi_{a+}\pi_{+a}}.$$

9 LASSO

- In the general case, the LASSO solution is the solution to $\text{argmin}_\beta - \ell(\beta) + \lambda \|\beta\|_1$, where ℓ is the log-likelihood function of interest.
- In our case, the LASSO is the solution to $\text{argmin}_\beta \frac{1}{2}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1$.
- The minimize $\hat{\beta}$ need not be unique, but the fitted value $X\hat{\beta}$ is always unique.
- Any 2 LASSO solutions must have the same signs in the overlap of their supports.
- **Equicorrelation set** $E := \{j \in \{1, \dots, p\} : |X_j^T(Y - X\hat{\beta})| = \lambda\}$. These variables achieve maximum inner product with the residual vector.
- The solution path $\lambda \mapsto \hat{\beta}(\lambda)$ is a piecewise, continuous function of λ with knots $\lambda_1 \geq \ldots \geq \lambda_r \geq 0$. The knots correspond to the values of λ at which the active set changes.
- **Model selection consistency:** See Lec 22 for details.
- The LASSO is most successful for sparse problems with features that are not too correlated.

9.1 Solving the LASSO

- **Coordinate descent:** Because the penalty is separable, we can solve the minimization problem by successively solving low-dimensional problems.
- **Proximal gradient descent**
- **Strong rules:** Say we are currently at a solution $\hat{\beta}_{\lambda_0}$. When fitting the LASSO at λ, we only include new variable j if $|X_j^T(Y - \hat{\beta}_{\lambda_0})| > 2\lambda - \lambda_0$.

9.2 KKT conditions for LASSO

- The **KKT conditions** are $X^T(Y - X\hat{\beta}) = \hat{u}$, where $\hat{u}$ is in the subgradient of the ℓ_1 norm evaluated at $\hat{\beta}$.
- The above can be rewritten as $X^T(Y - X\hat{\beta}) = \lambda s$, where
$$s_j \in \begin{cases} \{+1\} & \text{if } \hat{\beta}_j > 0, \\ \{-1\} & \text{if } \hat{\beta}_j < 0, \\ [-1, 1] & \text{if } \hat{\beta}_j = 0. \end{cases}$$

- For the equicorrelation variables, we can write the KKT conditions as $X_E^T(Y - X_E\hat{\beta}_E) = \lambda s_E$. This is a linear system that we can solve for $\hat{\beta}_E$.
- We thus get the representation

  $$\begin{cases} \hat{\beta}_E & = (X_E^T X_E)^\dagger (X_E^T Y - \lambda s_E) + \eta, \\ \hat{\beta}_{-E} & = 0, \end{cases}$$

  where the dagger indicates the pseudoinverse, and η is a vector in the null space of X_E.

  If X_E has full column rank, the LASSO solution is unique.
- Essentially, the LASSO constructs a map $X^T Y \mapsto (\hat{\beta}, \hat{u}) \in \{(\beta, u) : \beta \in N_u(K)\}$.

9.3 Group LASSO

- Let G be a partition of $\{1, \dots, p\}$. The group LASSO minimizes $\dfrac{1}{2}\|Y - X\beta\|_2^2 + \displaystyle\sum_{g \in G} \lambda_g \|\beta_g\|_2$.
- The group LASSO gives sparsity of groups instead of sparsity of variables.
- A common use of this is for categorical variables. If a variable has I levels, then we can group the I dummy variables together. (We put in all I variables instead of treating one as a baseline: otherwise, the LASSO will be biased toward the baseline.)

  For identifiability, the β's in each group must sum to 0.
- LASSO is the group LASSO with all groups of size 1.
- **KKT conditions:** For $\hat{\beta}_g \neq 0$, $u_g = \lambda_g \cdot \dfrac{\hat{\beta}_g}{\|\beta_g\|_2}$. For $\hat{\beta}_g = 0$, $\left\| \dfrac{u_g}{\lambda_g} \right\|_2 \leq 1$.
- (HW4)

  $$\text{argmin}_\beta \frac{L}{2}\|z - \beta\|_2^2 + \lambda\|\beta\|_2 = \frac{z}{\|z\|_2} \max\left(\|z\|_2 - \frac{\lambda}{L}, 0\right).$$

10 Kernel Methods

We observe $(X_1, Y_1), \dots (X_n, Y_n)$ with $X_i \in T$, $Y_i \in \mathbb{R}$ and we wish to solve

$$\underset{f \in \mathcal{C}}{\text{argmin}} \frac{1}{2}\|Y - f(X)\|_2^2,$$

where $\mathcal{C}$ is the span of a set of **flexible functions** $f_j : T \mapsto \mathbb{R}$, $j = 1, \dots, p$.

- A symmetric function $R : T \times T \mapsto \mathbb{R}$ is **non-negative definite** if for all k, $t_1, \dots, t_k \in T$ and $a \in \mathbb{R}^k$, we have $\displaystyle\sum_{i,j=1}^k a_i a_j R(t_i, t_j) = 0$.
- **Theorem:** Given a non-negative definite R, there exists a stochastic process $Z : \Omega \times T \mapsto \mathbb{R}$ such that $\text{Var}\left(\displaystyle\sum_{i=1}^k a_i Z_{t_i}\right) = \displaystyle\sum_{i,j=1}^k a_i a_j R(t_i, t_j)$. (In fact, we can take Z to be a Gaussian process.) R is called the **covariance function/kernel** of Z.

- From a covariance kernel R, we can get a Hilbert space

$$\mathcal{H} = \text{completion of } \left\{ \sum_i a_i R_{t_i} : \left\| \sum_i a_i R_{t_i} \right\|_{\mathcal{H}}^2 < \infty \right\},$$

  where $R_t : T \mapsto \mathbb{R}$ is defined by $R_t(s) = R(t, s)$, and the norm is from the inner product $\langle \sum_j a_j R_{t_j}, \sum_i a_i R_{s_i} \rangle_{\mathcal{H}} = \sum_{i,j} a_j b_i R(t_j, s_i)$.

- **Evaluation property:** For any $h \in \mathcal{H}$, we have $\langle h, R_t \rangle_{\mathcal{H}} = h(t)$.

Instead of the original minimization problem, consider the problem $\operatorname*{argmin}_{f \in \mathcal{H}} \frac{1}{2} \sum_{i=1}^n [Y_i - f(X_i)]^2 + \lambda \|f\|_{\mathcal{H}}^2$ instead.

Theorem: Consider the problem

$$\operatorname*{minimize}_{f \in \mathcal{H}} \quad \sum_{i=1}^n L\left(Y_i, f(X_i)\right) + \lambda \|f\|_{\mathcal{H}}^2,$$

where L is some loss function. If there is a minimizer, then all minimizers are in the linear span $\mathcal{L} = \text{span}(R_{X_1}, \ldots, R_{X_n})$.

- For covariance function R, define the **Gram matrix** G to be such that $G_{ij} = R(X_i, X_j)$.
- With the Gram matrix, we can rewrite the minimization problem as a finite dimensional problem of finding weights:

$$\operatorname*{minimize}_{w \in \mathbb{R}^n} \quad \frac{1}{2} \|Y - Gw\|^2 + \frac{\lambda}{2} w^T G w.$$

  Once we find solution, $\hat{w}$, we set $\hat{f} = \sum_{j=1}^n \hat{w}_j R_{X_j}$.

10.1 Kernel choices

- Gaussian kernel: $R(t, s) = \exp\left(-\frac{(t-s)^2}{2\gamma^2}\right)$ (infinitely smooth paths).
- Double exponential/Ornstein-Uhlenbeck kernel: $R(t, s) = \exp(-\gamma^{-1}|t - s|)$ (Hölder 1/2 paths).
- Brownian motion kernel: $R(t, s) = t \wedge s$ (Hölder 1/2 paths).
- **Linear kernel:** For $t \in \mathbb{R}^p$, define the Gaussian process $Z_t = \gamma^T t$, where $\gamma \sim \mathcal{N}(0, \Sigma)$. The corresponding covariance function is $R_t(s) = t^T \Sigma s$.
  - We find that $\mathcal{H} = \{\ell_a : \ell_a(x) = a^T x, a \in \text{row}(\Sigma)\}$, and $\langle \ell_a, \ell_b \rangle_{\mathcal{H}} = a^T \Sigma^\dagger b$.
- **Cubic smoothing splines:** $T = [0, 1]$. When we solve

$$\operatorname*{minimize}_{f} \frac{1}{2} \|Y - f(X)\|_2^2 + \frac{\lambda}{2} \int_{[0,1]} f''(t)^2 dt,$$

  where the minimization is over the class of twice-differentiable functions, the solution is a cubic spline.

  The covariance function corresponding to the cubic spline is $\text{Cov}(Z_s, Z_t) = \frac{ts(t \wedge s)}{2} - \frac{(t \wedge s)^3}{6}$.

- **Linear smoothing splines:** $T = [0,1]$. We seek to solve

  $$\underset{f}{\text{minimize}} \ \frac{1}{2}\|Y - f(X)\|_2^2 + \frac{\lambda}{2}\int_{[0,1]} f'(t)^2 dt.$$

  Here, the corresponding covariance function is $\text{Cov}(Z_t, Z_s) = t \wedge s$.

11 Survival Analysis

11.1 Set-up

We have observations (T_i, δ_i, X_i), where T_i is the failure/death time, δ_i is whether T_i is right-censored or not, and X_i are explanatory covariates.

Key objects of interest:

- **Survival function** $S(t) = \mathbb{P}(T > t) = 1 - F(t) = \bar{F}(t)$.
- **Hazard function** $h(t) = \lim_{\Delta \to 0} \dfrac{\mathbb{P}(t < T < t + \Delta \mid T \geq t)}{\Delta} = -\dfrac{S'(t)}{S(t)} = -\dfrac{d}{dt}\log S(t)$.
- If cumulative hazard $H(t) = \displaystyle\int_0^t h(s)ds$, then $S(t) = e^{-H(t)}$.

11.2 Survival analysis with 1 sample

Main goal is to estimate the survival function, i.e. get $\hat{S}(t)$ for different values of t.

- **Kaplan-Meier method:** Let $t_{(1)} < \cdots < t_{(m)}$ be the ordered times of death. Let d_i be the number of deaths at time $t_{(i)}$, and let n_i be the number alive just before $t_{(i)}$ (i.e. the number "at risk"). Then the Kaplan-Meier estimate of the survival function is

  $$\hat{S}(t) = \prod_{i:t_{(i)}<t}\left(1 - \frac{d_i}{n_i}\right).$$

  Justification: $S(t) = \mathbb{P}(T > t_{(1)})\mathbb{P}(T > t_{(2)} \mid T > t_{(1)})\ldots$, and $1 - \frac{d_i}{n_i}$ estimates $\mathbb{P}(T > t_{(i)} \mid T > t_{(i-1)})$.

To estimate the cumulative hazard function, we can use the **Nelson-Aalen method:**

- Under no censoring and assuming no ties in data, we have $\hat{H}(t) = \displaystyle\sum_{T_i \leq t} \frac{1}{\#\{T_j \geq t\}}$.
- Under right-censoring, using the notation for the Kaplan-Meier method, $\hat{H}(t) = \displaystyle\sum_{t_{(i)} \leq t} \frac{d_i}{n_i}$.

11.3 Survival analysis with 2 samples

The task is to test the null hypothesis that the 2 samples come from the same (survival) distribution. This can be accomplished by the **log-rank test**. (The log-rank test works for right-censored data.)

Let $t_{(1)} < \cdots < t_{(m)}$ be the ordered times of death for both samples. For each i, set-up a 2×2 contingency table, tabulating the number which survived and died from each sample at time $t_{(i)}$ (rows are survived/died, columns are which group the sample is from). If the 2 samples were the same, then there would be independence in each 2×2 table. Hence, we apply the Cochran-Mantel-Haenszel test stratified across time.

11.4 Survival regression

Strategies to model the effect of covariates on survival time:

- **Accelerated failure time:** Model $S(t) = S_0\left(e^{-x^T\beta}t\right)$, where S_0 is some baseline survival function.
- **Proportional hazards model:** Model $h(t) = h_0(t)e^{x^T\beta}$, where $h_0(t)$ is the baseline hazard.
- **Cox proportional hazards model:** As above, but we try to get away with not making any assumption on $h_0(t)$.

  Let $t_{(1)} < \cdots < t_{(m)}$ be the times of death. At time $t_{(i)}$, say individual $j(i)$ dies. We ask, what is the probability that particular individual died, given that someone died? Letting R_i be the risk set at time $t(i)$ (i.e. the people still alive just before time $t(i)$),

  $$\mathbb{P}(j(i) \text{ died} \mid \text{someone died}) = \frac{\exp\left(x_{j(i)}^T\beta\right)}{\sum_{j\in R_i}\exp\left(x_j^T\beta\right)}.$$

  Multiplying across i, we get Cox's **partial likelihood** $L(\beta) = \prod_{i=1}^{m}\frac{\exp\left(x_{j(i)}^T\beta\right)}{\sum_{j\in R_i}\exp\left(x_j^T\beta\right)}$. Cox showed that we can treat this like a real likelihood and carry out MLE estimation, etc.

12 EM Algorithm

The EM algorithm is an iterative approach to maximizing a likelihood, designed for the case when there are latent variables or missing data in the problem. We might use this if

- There is no closed form for the MLE, or
- The log-likelihood function is non-convex.

Assume that we have data X and latent variables Z jointly following the law $L(\theta; X, Z) = p_\theta(X, Z)$. We do not observe latent variables Z, so we must work with

$$L(\theta; X) = p_\theta(x) = \int_z p_\theta(x, z)dz.$$

While we do not see the complete data log-likelihood, we can average over the z's to get the expected complete data log-likelihood.

Expectation (E) step: Calculate the expected value of the log-likelihood function w.r.t. the conditional distribution of Z given X, under the current estimate $\hat{\theta}^k$ for θ:

$$Q\left(\theta \mid \hat{\theta}^k\right) := \mathbb{E}_{Z|X,\hat{\theta}^k}\left[\log L(\theta; X, Z)\right].$$

Maximization (M) step: Find the value of θ that maximizes the quantity above and let that be the new estimate, i.e.

$$\hat{\theta}^{k+1} := \underset{\theta}{arg\,max} Q\left(\theta \mid \hat{\theta}^k\right).$$

Guarantee: EM will increase $\ell(\theta)$ (the value of the log-likelihood) at each iteration, which implies that it will converge to a local maximum. (See Gene's notes for an explanation for why the EM algorithm works.)

Gene's version of EM

Assume that we have parameters θ which we want to estimate. These parameters generate observed variables X, and unobserved variables Z. Assume that given θ, Z and X have a joint density $p_\theta(X, Z)$.

If we could observe Z, we have a **complete data log-likelihood** $\ell(\theta; X, Z) = \log p_\theta(X, Z)$. Unfortunately, we can't observe Z. Instead, assume we have some initial guess $\theta^{(k)}$ for θ.

Expectation (E) step: Calculate the expected value of the log-likelihood function w.r.t. the conditional distribution of Z given X, under the current estimate $\hat{\theta}^{(k)}$ for θ:

$$\hat{\pi}^{(k)} := \mathbb{E}_{\hat{\theta}^{(k)}}\left[Z \mid X\right].$$

Then write out the **expected complete data log-likelihood**, i.e. basically plugging in $\hat{\pi}^{(k)}$ for Z in the complete data log-likelihood:

$$\tilde{\ell}^{(k)}(\theta; X) := \mathbb{E}_{\hat{\theta}^{(k)}}\left[\ell(\theta; X, Z) \mid X\right].$$

Maximization (M) step: $\tilde{\ell}^{(k)}(\theta; X)$ is a function of θ. Maximize it to get $\hat{\pi}^{(k+1)}$.

13 Lindsey's Method

Suppose we wish to estimate the probability density $g(y)$ that produced observed random sample $Y_i \overset{iid}{\sim} g(y)$, $i = 1, \ldots, n$.

Say we have a histogram data for the y_i's, i.e. we have bins $1, \ldots, K$, which are centered at $x_1, \ldots, x_K$, and we have the counts of the number of y_i's which fall into each bin, $Z_1, \ldots, Z_K$. Then $(Z_1, \ldots, Z_K) \sim \text{Multinom}(n, (p_1, \ldots, p_K))$, where p_i is the probability of Y falling into bin i.

The trick: Instead of assuming n is known, assume that $n \sim$ Pois(μ) for unknown μ. With this assumption, $Z_j \overset{ind.}{\sim} \text{Pois}(\mu p_j)$. Let $\mu_j = \mu p_j$.

If we have a parametric form for g, we can write each p_j in terms of x_j. We will usually end up getting $p_j = e^{\beta_0 + \beta_1 f_1(x_j) + \cdots + \beta_p f_p(x_j)}$ for some functions $f_1, \ldots, f_p$.

We can then do standard Poisson regression: $Z_j \overset{ind.}{\sim} \text{Pois}(\mu_j)$, $\log \mu_j = \beta_0 + \beta_1 f_1(x_j) + \cdots + \beta_p f_p(x_j)$.