

## Homework # 4

Due in Class Tuesday, December 10

1. Consider the S-plus air data set you used in HW2. Consider the nonparametric model

$$Y_i = \theta_1(X_{1i}) + \theta_2(X_{2i}) + \epsilon_i, \quad (1)$$

where  $Y=OZONE$ ,  $X_1=WIND$ ,  $X_2=TEMPERATURE$ ,  $\epsilon_i \sim N(0, \sigma^2)$ .

(1) Fit this model using loess to estimate  $\theta_1(X_1)$  and  $\theta_2(X_2)$ . Construct 95% CIs of the curves. You could use the S-plus function `plot(gam-object, se=T, residuals=T)` to get CIs and scatter plots.

(2) Fit this model using spline smoothing to estimate  $\theta_1(X_1)$  and  $\theta_2(X_2)$ . Construct 95% CIs of the curves.

(3) Use the SAS macro GAMM1.MAC, which can be downloaded from the class website, to fit a mixed model to obtain smoothing spline estimates of  $\theta_1(X_1)$  and  $\theta_2(X_2)$  in model (1). Plot the estimated curves and their 95% frequentist and Bayesian CIs and comment.

(4) Use PROC MIXED to fit (1) using smoothing splines and compare your results with those obtained from GAMM1.MAC. Use PROC MIXED to construct 95% Bayesian CIs of the curves.

2. Consider the respiratory infection data “indon\_base.dat”, which contains the baseline data of 275 Indonesian children including their respiratory infection status and covariates. The variables are ID, XERO (vitamin A deficiency status (Y/N)), AGE (centered by 36 months), SEX, HEIGHT and INFECT (Y/N). Consider the nonparametric logistic model

$$\text{logit}\{E(Y_i = 1)\} = \theta(X_i),$$

where  $Y=INFECT$  and  $X=AGE$ .

(1) Estimate  $\theta(x)$  using a smoothing spline via the Splus function GAM.

(2) Write your own code to fit this model by iteratively fitting a weighted smoothing spline under

$$Z_i = \theta(X_i) + \epsilon_i, \quad (2)$$

where  $Z_i$  is the the GNM working vector and  $\epsilon_i \sim N(0, w_i^{-1})$ , and  $w_i$  is the GNM working weight matrix. See the GNM smoothing spline lecture notes for more details. At each iteration, you could fit (2) using the Splus function `smooth.spline`. For simplicity, you could use the GAM estimate of the smoothing parameter  $\lambda$ . Compare your results with those from GAM.

3. Consider GEEs for analysis of longitudinal/clustered data. Suppose the data consist of  $n$  subjects with  $m_i$  observations per subject with an outcome variable  $Y_{ij}$  and a  $p \times 1$  vector of covariates  $\mathbf{X}_{ij}$ , where  $i$  indexes subject  $i$  and  $j$  indexes observation  $j$ . Suppose the marginal mean and variance of  $Y_{ij}$  are  $E(Y_{ij}|\mathbf{X}_i) = \mu_{ij}$  and  $\text{var}(Y_{ij}|\mathbf{X}_i) = \phi a_{ij}^{-1} v(\mu_{ij})$ , where  $v(\cdot)$  is a variance function. Consider the marginal GLM

$$g(\mu_{ij}) = \mathbf{X}_{ij}^T \boldsymbol{\beta},$$

where  $g(\cdot)$  is a link function and  $\boldsymbol{\beta}$  is a  $p \times 1$  vector of regression coefficients. Consider the use of GEEs to estimate  $\boldsymbol{\beta}$ ,

$$\sum_{i=1}^n \mathbf{D}_i^T \mathbf{V}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) = 0,$$

where  $\mathbf{D}_i$  and  $\mathbf{V}_i$  are defined in class.

(1) Read Sections 9.1-9.5 of McCullagh and Nelder (*Generalized Linear Models*, Chapman and Hall, 1989), Liang and Zeger (1986, *Biometrika*, 73, 13-22) and Foutz (1997, *JASA*, 72, 147-148). You can download the papers from the website <http://www.jstor.org/cgi-bin/jstor/listjournal>. Show that the GEE estimator  $\hat{\boldsymbol{\beta}}$  is consistent and asymptotically normal for any working correlation matrix  $\mathbf{R}_i$ .

(2) Show that when the working correlation matrix  $\mathbf{R}_i$  is equal to the true correlation, the GEE estimator  $\hat{\boldsymbol{\beta}}$  is most efficient within the linear estimating equation family.

(3) Consider the model

$$g(\mu_{ij}) = \mathbf{X}_{1ij}^T \boldsymbol{\beta}_1 + \mathbf{X}_{2ij}^T \boldsymbol{\beta}_2.$$

Derive the score test for testing for  $H_0 : \boldsymbol{\beta}_2 = 0$  against  $H_1 : \boldsymbol{\beta}_2 \neq 0$ . See the lecture notes for details.

4. Consider the infectious disease data discussed in class. You can download it from the class website "indon.dat". Fit a logistic model using covariates VISIT, SEX, XERO, HEIGHT, STUNTING, and AGE, where AGE is the current age.

(1) Fit a parametric logistic model using covariates COS(VISIT), SIN(VISIT), SEX, XERO, HEIGHT, STUNTING, and linear and quadratic AGE using GEEs. Is the quadratic age effect significant? How about the cubic age effect? Try a few different working correlation matrices and compare results. Interpret the results.

(2) Fit a parametric logistic model using covariates SEX, XERO, HEIGHT, STUNTING, dummy variables for VISIT, BASELINE AGE (linear and quadratic) using GEE. Try a few different working correlation matrices and compare results. Interpret the results. Compare the results in (1) and discuss interpretation differences.

(3) Consider a nonparametric model

$$\text{logit}\{E(Y_{ij} = 1)\} = \theta(\text{AGE}_{ij}),$$

where  $\theta(\cdot)$  is an unknown smooth function. Fit this model using the local linear kernel method assuming working independence and construct the 95% CI of the curve using a sandwich estimator. Plot the estimated curve and its 95% CI.

(5) Compare the nonparametric fit with the quadratic fit of the age effect under  $\text{logit}\{E(Y_{ij} = 1)\} = \beta_0 + \beta_1 \text{AGE} + \beta_2 \text{AGE}^2$  and comment.