# Exam Statistical Models

18 December 2013

*Please write your name and student number on each page you turn in. Motivate your answers. Write your solution clearly, using consistent notation. You may use a simple calculator provided it is not part of a device that is capable of communication with other devices.*

1. The moisture content of three types of cheese made by two methods was recorded. Two pieces of cheese were measured for each type and each method: $Y_{ijk}$ denotes the moisture content in $k$th piece of cheese for method $i$ and cheese type $j$, $i = 1, 2$, $j = 1, 2, 3$ and $k = 1, 2$.

   (i) (7 pts.) Write the appropriate two-way ANOVA model that can be applied to investigate the effects of cheese type and method (and their interaction) on the moisture content. Specify the design matrix, all model assumptions and the constraints needed to make the model identifiable.

   (ii) (10 pts.) After fitting the ANOVA model to the data, an ANOVA table is obtained. This table is partially presented below. Provide the missing information.

| Source | Sum of Squares | Df | Mean Square | $F$-statistic | $p$-value |
|---|---|---|---|---|---|
| Method | 0.1141 | ___ | _______ | _______ | 0.3485 |
| Type | ________ | 2 | 12.9501 | ________ | 0.0000155 |
| Interaction | 0.3026 | ___ | _______ | 1.37126 | 0.3233 |
| Residuals | 0.6620 | ___ | 0.1103 | | |
| Total | ________ | ___ | | | |

   (iii) (7 pts.) Let the significance level $\alpha = 0.01$. Based on the ANOVA table in part (ii), carry out a two-way ANOVA for the both factors (Type and Method) and their interaction.

   (vi) (8 pts.) Suppose that, based on the ANOVA table in part (ii), one decides to fit a one-way ANOVA model instead. Which factor is then to use? Present schematically the corresponding one-way ANOVA table (without specifying the numbers in it), provide only the numbers in the column "Df" (degrees of freedom).

2. Suppose we have a dataset $\{(Y_1, x_1), \dots, (Y_n, x_n)\}$ which is modeled as follows:

$$Y_i = f(x_i, \boldsymbol{\theta}) + \varepsilon_i, \quad i = 1, \dots, n, \qquad (*)$$

   where $\boldsymbol{\theta} = (\theta_1, \theta_2)$ is to be estimated and such that $\theta_1 \neq 0$, $f(x_i, \boldsymbol{\theta}) = \sin(\theta_1 x_i) + \theta_1 \exp\{-\theta_2 x_i\}$, $\varepsilon_1, \dots \varepsilon_n$ are independent random errors such that $\mathbb{E}\varepsilon_i = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$, $i = 1, \dots, n$.

   (i) (6 pts.) Suppose $n = 200$, $x_1 = x_2 = \dots = x_{100} = 0$ and $x_{101} = x_{102} = \dots = x_{200} = 1$. Propose a starting value for the LSE $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2)$ in the Gauss-Newton method and explain your choice.

   (ii) (6 pts.) The normal equations (used for calculating the LSE of $\boldsymbol{\theta}$) are $\sum_{i=1}^n \frac{\partial f}{\partial \theta_l}(x_i, \boldsymbol{\theta})(Y_i - f(x_i, \boldsymbol{\theta})) = 0$, $l = 1, \dots, p$. Give the normal equations for the model $(*)$.

   (iii) (6 pts.) Suppose we obtained the LSE $\hat{\boldsymbol{\theta}} = (2.28, 1.52)$ for the parameter $\boldsymbol{\theta}$ and an estimator for the covariance matrix of $\hat{\boldsymbol{\theta}}$

$$\widehat{\mathrm{Cov}(\hat{\boldsymbol{\theta}})} = \hat{\sigma}^2 (\hat{V}^T \hat{V})^{-1} = \begin{pmatrix} 1.23 & 0.43 \\ 0.43 & 0.64 \end{pmatrix}.$$

Use the above matrix and the quantile $t_{198;0.975} = 1.972$ to construct a 95% (approximate) confidence interval for $\theta_2$ and to test the hypothesis $H_0 : \theta_2 = 0$.

3. Suppose $n$ independent trials are performed. At the $i$-th trial we observe $Z_i \sim \text{Bin}(5, \pi_i)$, $\pi_i \in (0, 1)$, $i = 1, \ldots, n$, i.e.,

$$P(Z_i = k) = \binom{k}{5} \pi_i^k (1 - \pi_i)^{5-k}, \quad k = 0, 1, \ldots, 5.$$

Besides, at each trial the values of two covariates are available, called, say, covariate A and covariate B. We use a logistic regression model with two covariates.

(i) (7 pts.) Write down the model, including the assumptions.

(ii) (8 pts.) The general form of the exponential family is

$$f(y, \theta_i) = \exp\left\{\frac{y\theta_i - b(\theta_i)}{\phi/A_i} + c(y, \phi/A_i)\right\}.$$

Show that the distribution of $Z_i$ can be written in this form with parameter $\theta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right)$. Identify the function $b(\theta)$ and the parameters $\phi$ and $A_i$.

(iii) (10 pts.) Suppose we obtained the following analysis of deviance table.

| Terms | Resid. Df | Residual Deviance | Test | Df | Deviance reduction |
|---|---|---|---|---|---|
| B + A + I | 50 | 40.45 | | | |
| A + I | 51 | 42.34 | -B | -1 | -1.89 |
| I (Intercept) | 52 | 47.49 | -A | -1 | -5.15 |

Fix the significance level $\alpha = 0.05$. What can you tell about the relevance of the covariates A and B in the model? Does the full model fit well? You can use the following facts: we reject $H_0$ : `the model fits well` if $D/\phi > \chi^2_{n-p-1;1-\alpha}$, where $D$ is the residual deviance for the fitted model; $\chi^2_{1;0.95} = 3.84$, $\chi^2_{2;0.95} = 5.99$ and $\chi^2_{50;0.95} = 67.5$; $\phi = 1$ for the binomial model.

4. Let $\{Z_t\}$ denote a white noise time series with variance $\sigma^2$.

(i) (8 pts.) Show that the time series $\{X_t\}$ given by $X_t = tZ_t + t^2$ is *not* weakly stationary. Let $Y_0 = Z_0$ and $Y_t = (X_t - t^2)/t$ for $t \neq 0$. Is $\{Y_t\}$ weakly stationary?

(ii) (9 pts.) Let $\{X_t\}$ be the MA(2) time series given by

$$X_t = Z_t + 3Z_{t-2}.$$

Compute $\gamma_X(0), \gamma_X(1), \gamma_X(2)$. Consider the time series $\{Y_t\}$ given by $Y_t = \nabla X_t$. Show that $\{Y_t\}$ is a MA(3) time series, and identify the values of the coefficients $\beta_1, \beta_2, \beta_3$.

(iii) (8 pts.) Consider a stationary (i.e., with $|\alpha| < 1$) AR(1) time series:

$$X_t = \alpha X_{t-1} + Z_t.$$

Derive the Yule-Walker equations for this model and argue how these can be used to estimate $\alpha$ and $\sigma^2$.