Practice exercises for exam Statistical Models

*Motivate your answers. Write your solution clearly, using consistent notation. You may use a simple calculator provided it is not part of a communicating device. Regular question form the practice exam, questions marked with * are additional. You can use the following quantiles:* $F_{1,6,0.99} = 13.75$, $F_{4,10,0.995} = 7.34$, $\chi^2_{1;0.95} = 3.84$, $\chi^2_{2;0.95} = 5.99$, $\chi^2_{3;0.95} = 7.81$, $t_{198;0.975} = 1.972$, $\chi^2_{50;0.95} = 67.5$, $\chi^2_{52;0.95} = 69.83$.

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 kth piece of cheese for method i and cheese type j, $i = 1, 2$, $j = 1, 2, 3$ and $k = 1, 2$. The data is summarized in the following table of moisture averages for each cheese type and each method.

	Type 1	Type 2	Type 3	row average
Method 1	38.905	35.575	36.51	36.99667
Method 2	38.985	35.55	35.87	36.80167
column average	38.945	35.5625	36.19	36.89917

   (i) (7 p.) 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)* (5 p.) Under the model you specified in part (i), use the information in the above table to determine the least squares estimates of the main effect corresponding to Type 1 and the interaction effect between Type 2 and Method 2.

   (ii) (9 p.) After fitting the ANOVA model to the data, an ANOVA table is obtained. This table is partially presented below. Provide the missing information (where possible).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Method	___	0.1141	_____	_____	
Type	2	_____	12.9501	_____	0.0000155
Interaction	___	0.3026	_____	1.3717	0.3233
Residuals	___	0.6620	0.1103		

   (iii) (8 p.) 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. Is it justified to reject the full model in favor of the additive model?

   (vi) (8 p.) 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`.

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

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

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

   (i) (7 p.) Suppose $n = 200$, $x_1 = x_2 = \ldots = x_{100} = 0$ and $x_{101} = x_{102} = \ldots = x_{200} = 1$. 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) (4 p.) Give the normal equations (used for calculating the LSE of $\boldsymbol{\theta}$) for the model $(*)$.

   (iii) (7 p.) 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 $\hat{\boldsymbol{\theta}}$

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

   Construct a 95% (approximate) confidence interval for θ_2 and test the hypothesis $H_0 : \theta_2 = 0$.

   (iii)* (5 p.) Suppose that after plotting the curve corresponding to $f(x, \hat{\boldsymbol{\theta}})$ using the LSE $\hat{\boldsymbol{\theta}}$ obtained from the Gauss-Newton method, the researcher observes that the curve does not fit the data well. State one likely reason for the bad fit.

(iv)* (5 p.) Suppose the researcher fits this non-linear model to a set of data. Based on 1000 samples of the centered residuals, he computes 1000 bootstrap estimates θ_2^* of θ_2. Let $\theta_{2;q}^*$ denote the bootstrap estimate of θ_2 that is in position q when the bootstrap estimates are ordered from smallest to largest. If $\theta_{2;26}^* = -0.07$ and $\theta_{2;976}^* = 3.1$, what is the 95% bootstrap confidence interval for θ_2?

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.,

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

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 p.) Write down the model, including the assumptions.

(ii) (9 p.) The general form of the exponential family is

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

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

(iii) (9 p.) Suppose we obtained the following (partial) analysis of deviance table.

	Df	Deviance	Resid. Df	Resid. Dev	P(> |Chi|)
NULL			52	47.49	
A	1	5.15	51	42.34	
B	1	1.89	50	40.45	

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?

(iii)* (5 p.) Based on the table in (iii), which covariate has the least significant effect on the response variable? Suppose the Pearson chi-squared statistic for this fit is $P = 43.76$, what is the value of the estimate for ϕ?

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

(i) (7 p.) Is the time series $\{X_t\}$ given by $X_t = tZ_t + t^2$ weakly stationary? Let $Y_0 = Z_0$, $Y_t = (X_t - t^2)/t$ for $t \neq 0$ and $V_t = \nabla^2 X_t$. Are $\{Y_t\}$ and $\{V_t\}$ weakly stationary?

(i)* (8 p.) Are the time series $\{X_t\}$ and $\{Y_t\}$ given by $X_t = Z_0 \cos(t) + Z_2 \sin(t)$ and $Y_t = Z_t \cos(t) + Z_{t+2} \sin(t)$ weakly stationary? (Hint: use $\cos(\alpha+\beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta$.)

(ii) (9 p.) 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(q) time series, and identify the values of the parameter q and the coefficients $\beta_1, \ldots, \beta_q$.

(ii)* (7 p.) For the time series $\{X_t\}$ from (ii), consider $Y_t = X_t + a + S_t + bt$ for some $a, b \in \mathbb{R}$ and some seasonal component S_t with period $d \in \mathbb{N}$. Propose a linear filter for $\{Y_t\}$ leading to a stationary time series. Is this resulting time series of ARMA type? If so, what type?

(iii) (9 p.) Consider a stationary (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 α and σ^2.

(iii)* (6 p.) Assuming stationarity for the time series $\{X_t\}$ from (iii), find $\mathbb{E}X_t$, $\text{Var}(X_t) = \gamma_X(0)$, $\gamma_X(1)$, $\gamma_X(2)$, $\rho(1)$ and $\rho_X(2)$.