Short solutions for practice exam Statistical Models

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. 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.

   Solution. $Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$, $e_{ijk} \overset{\text{ind}}{\sim} N(0, \sigma^2)$, $i = 1, 2$, $j = 1, 2, 3$, $k = 1, 2$. $\sum_{i=1}^{2} \alpha_i = 0$; $\sum_{j=1}^{3} \beta_j = 0$; $\sum_{i=1}^{2} \gamma_{ij} = 0$, $j = 1, 2, 3$; $\sum_{j=1}^{3} \gamma_{ij} = 0$, $i = 1, 2$. The first column of the design (12×12)-matrix X exists of 1's, the second of six 1's and six 0's etc.

   (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	1	0.1141	0.1141	1.0345	
Type	2	25.9002	12.9501	117.408	0.0000155
Interaction	2	0.3026	0.1513	1.3717	0.3233
Residuals	6	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?

   Solution. Let $H_A : \alpha_i = 0$, $i = 1, 2$ (method has no effect), $H_B : \beta_j = 0$, $j = 1, 2, 3$ (type has no effect), $H_{AB} : \gamma_{ij} = 0$, $i = 1, 2$, $j = 1, 2, 3$ (no interaction between Method and Type). By looking at the last column of the above table, at $\alpha = 0.01$, we do not reject the hypothesis H_{AB}, but reject H_B. Hence, it is not justified to reject the full model in favor of the additive model. As to the hypothesis H_A, compare the value of F-statistic $f = 1.0345$ with the corresponding F-quantile (namely, $F_{1,6,0.99} = 13.75$): since $f > F_{1,6,0.99}$ is not true, H_A is not rejected, i.e., factor Method has no effect.

   (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`.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Type	2	*	*	*	*
Residuals	9	*	*		

   Remark. Notice that, although this is not asked, many other entries in the above one-way ANOVA table can actually be determined by using the two-way ANOVA table from (ii). Try to do this by yourself as this can be asked at the exam.

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.
Solution. Put $x_1 = x_2 = \ldots = x_{100} = 0$ in $(*)$ so that we have $Y_i = \theta_1 + \varepsilon_i$, $i = 1, \ldots, 100$. Clearly $\tilde{\theta}_1 = \frac{1}{100}\sum_{i=1}^{100} Y_i$ is a good estimator of θ_1. Putting now $x_{101} = x_{102} = \ldots = x_{200} = 1$ in $(*)$, we obtain $Y_j = \sin(\theta_1) + \theta_1 e^{-\theta_2} + \varepsilon_j$, $j = 101, \ldots, 200$, so that $T = \frac{1}{100}\sum_{j=101}^{200} Y_j$ is a good estimator of $\sin(\theta_1) + \theta_1 e^{-\theta_2}$. Then $\tilde{\theta}_2 = -\log\left(\frac{T - \sin(\tilde{\theta}_1)}{\tilde{\theta}_1}\right)$ is a good estimator of θ_2. We take $(\tilde{\theta}_1, \tilde{\theta}_2)$ as a starting value.

(ii) (4 p.) Give the normal equations (used for calculating the LSE of $\boldsymbol{\theta}$) for the model $(*)$.
Solution. $\sum_{i=1}^{n} \left(x_i \cos(\theta_1 x_i) + e^{-\theta_2 x_i}\right)\left(Y_i - \sin(\theta_1 x_i) - \theta_1 e^{-\theta_2 x_i}\right) = 0$,
$\sum_{i=1}^{n} x_i e^{-\theta_2 x_i}\left(Y_i - \sin(\theta_1 x_i) - \theta_1 e^{-\theta_2 x_i}\right) = 0$.

(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 $H_0 : \theta_2 = 0$.
Solution. Since $\hat{\theta}_l \pm t_{n-p;1-\alpha/2}\hat{\sigma}\sqrt{((\hat{V}^T\hat{V})^{-1})_{ll}}$ is a $(1-\alpha)$-confidence interval for θ_l, $l = 1, \ldots, p$, in our case we obtain that $\hat{\theta}_2 \pm t_{198;0.975}\sqrt{0.64} = 1.52 \pm 1.972 \cdot 0.8 = (-0.0576, 3.0976)$ is an approximate 0.95-confidence interval. The $H_0 : \theta_2 = 0$ is not rejected since 0 lies in that confidence interval.

3. Suppose n independent trials are performed. At the i-th trial we observe $Z_i \sim \mathrm{Bin}(5, \pi_i)$, $\pi_i \in (0, 1)$, $i = 1, \ldots, n$, 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.
Solution. Logistic regression model: $Z_i \sim \mathrm{Bin}(5, \pi_i)$, $\pi_i \in (0, 1)$, $\eta_i = x_{iA}\beta_1 + x_{iB}\beta_2$, $\eta_i = g(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right)$, $i = 1, \ldots, n$.

(ii) (9 p.) 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 ϕ and A_i.
Solution. Let $\theta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right)$, then $\pi_i = \frac{e^{\theta_i}}{1+e^{\theta_i}}$. Write $P(Z_i = y) = \binom{5}{y}\pi_i^y(1-\pi_i)^{5-y} = \exp\left\{y\log\left(\frac{\pi_i}{1-\pi_i}\right) + 5\log(1-\pi_i) + \log\left(\frac{5!}{y!(5-y)!}\right)\right\} = \exp\left\{\frac{y\theta_i - b(\theta_i)}{\phi/A_i} + c(y, \phi/A_i)\right\}$, where $b(\theta) = -5\log(1-\pi_i) = -5\log\left(1 - \frac{e^{\theta_i}}{1+e^{\theta_i}}\right) = 5\log(1+e^{\theta_i})$. Further, $\phi = 1$, $A_i = 1$ and $c(y, \phi/A_i) = \log\left(\frac{5!}{y!(5-y)!}\right)$.

(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?

Solution: Since the deviance reduction for covariate A $5.15 > \chi^2_{1;0.95} = 3.84$, covariate A is relevant. On the other hand, the deviance reduction for covariate B $1.89 < \chi^2_{1;0.95}$, covariate B is not relevant. Also if we compare the largest model (I+A+B) with the smallest (I), we see that the hypothesis for the smallest model would have been rejected since $5.15+1.89 > \chi^2_{2;0.95} = 5.99$. So the covariates A and B together are relevant as compared to the smallest model.

Next, since for the full model $D/\phi = 40.45/1 = 40.45 < \chi^2_{50;0.95} = 67.5$, we do not reject H_0 : `the model fits well`. This means that the full model fits well.

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 $Y_0 = Z_0$ and $Y_t = (X_t - t^2)/t$ for $t \neq 0$. Is $\{Y_t\}$ weakly stationary?

Solution: Since $\mathrm{E}X_t = t^2$ (not constant), $\{X_t\}$ is not weakly stationary. Notice that $Y_t = Z_t$, white noise time series which is known to be weakly stationary. Thus $\{Y_t\}$ is weakly stationary.

(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 $\beta_1, \ldots, \beta_q$.

Solution: Compute $\gamma_X(0) = \mathrm{E}X_t^2 = \mathrm{E}Z_t^2 + 9\mathrm{E}Z_{t-2}^2 = 10\sigma^2$, $\gamma_X(1) = \mathrm{E}(X_tX_{t+1}) = 0$ and $\gamma_X(2) = \mathrm{E}(X_tX_{t+2}) = \mathrm{E}\big((Z_t + 3Z_{t-2})(Z_{t+2} + 3Z_t)\big) = 3\mathrm{E}Z_t^2 = 3\sigma^2$. Next, $Y_t = \nabla X_t = X_t - X_{t-1} = Z_t + 3Z_{t-2} - Z_{t-1} - 3Z_{t-3} = Z_t - Z_{t-1} + 3Z_{t-2} - 3Z_{t-3}$ so that $\{Y_t\}$ is a MA(3) time series with $\beta_1 = -1$, $\beta_2 = 3$, $\beta_3 = -3$.

(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.

Solution: Since $\mathrm{E}X_t = 0$ and X_{t-1} and Z_t are uncorrelated, $\gamma(0) = \mathrm{E}(X_tX_t) = \mathrm{E}(\alpha^2X_{t-1}^2) + \mathrm{E}(Z_t^2) = \alpha^2\gamma(0) + \sigma^2$, which implies the known result $\gamma(0) = \frac{\sigma^2}{1-\alpha^2}$. Similarly, $\gamma(1) = \mathrm{E}(X_tX_{t-1}) = \mathrm{E}\big((\alpha X_{t-1} + Z_t)X_{t-1}\big) = \alpha\gamma(0)$. Thus $\alpha = \frac{\gamma(1)}{\gamma(0)}$ and $\sigma^2 = (1-\alpha^2)\gamma(0) = \gamma(0) - \frac{\gamma^2(1)}{\gamma(0)}$. Therefore we can estimate the parameters α and σ^2 as follows:

$$\hat{\alpha} = \frac{\hat{\gamma}(1)}{\hat{\gamma}(0)}, \qquad \hat{\sigma}^2 = \hat{\gamma}(0) - \frac{\hat{\gamma}^2(1)}{\hat{\gamma}(0)}.$$

Here $\hat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-h}(X_{t-h} - \bar{X}_n)(X_t - \bar{X}_n)$, $h \geq 0$, is the sample autocovariance function and $\bar{X}_n = \frac{1}{n}\sum_{t=1}^{n} X_t$ is the sample mean.