

Family Name
First Name
Student Number

UNIVERSITY OF WOLLONGONG
SCHOOL OF MATHEMATICS AND APPLIED STATISTICS

STAT902 : Advanced Data Analysis

Autumn Session Examination 2008

Time Allowed: 3 hours 15 minutes
Number of Questions: 6

DIRECTIONS TO CANDIDATES

1. Each question is to be attempted.
2. All questions are of equal value.
3. The examination paper is printed on both sides.
4. A list of possibly useful facts is provided at the end of the examination paper.

EXAMINATION MATERIALS/AIDS ALLOWED

Pens
Ruler

THIS EXAMINATION PAPER MUST NOT BE REMOVED FROM THE EXAMINATION ROOM
--

1. (a) According to a particular model, the exchange rates, X and Y , of two currencies with respect to the Australian dollar have joint distribution

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N \left(\begin{bmatrix} 41 \\ 26 \end{bmatrix}, \begin{bmatrix} 9 & 8 \\ 8 & 25 \end{bmatrix} \right)$$

If X is observed to be 39.2 then what is the best prediction of Y ?

- (b) Let X_1, \dots, X_n be modelled as a random sample from the distribution

$$[x; \alpha] = \alpha x^{\alpha-1} e^{-x^\alpha}, \quad x > 0, \alpha > 0.$$

Write an algorithm for maximum likelihood estimation of α . The algorithm should be refined to the point that it can be directly implemented in a computer programming language such as R.

2. (a) Consider the generalised linear model with log-likelihood

$$\ell(\boldsymbol{\beta}) = \mathbf{y}^T \mathbf{X}\boldsymbol{\beta} + \mathbf{1}^T (-2\mathbf{X}\boldsymbol{\beta})^{1/2} + \mathbf{1}^T c(\mathbf{y}). \quad (1)$$

Here \mathbf{y} is a vector of responses, \mathbf{X} is a design matrix of predictors and $c(\mathbf{y})$ involves normalising factors that do not depend on $\boldsymbol{\beta}$.

- (i) Derive the Fisher information matrix of $\boldsymbol{\beta}$.
(ii) A quasi-likelihood extension of (1) is

$$q\ell(\boldsymbol{\beta}, \phi) = (\mathbf{y}^T \mathbf{X}\boldsymbol{\beta} + \mathbf{1}^T (-2\mathbf{X}\boldsymbol{\beta})^{1/2})/\phi$$

where $\phi > 0$ is a dispersion parameter. If $\hat{\boldsymbol{\beta}}$ is the estimate of $\boldsymbol{\beta}$ and ϕ is estimated to be 4.71 then obtain an expression for the standard errors of the entries of $\hat{\boldsymbol{\beta}}$.

- (b) The simple random intercept model for longitudinal data is

$$y_{ij} = \beta_0 + U_i + \beta_1 x_{ij} + \varepsilon_{ij}.$$

where

$$U_i \stackrel{\text{ind.}}{\sim} N(0, \sigma_U^2), \quad \text{independently of} \quad \varepsilon_{ij} \stackrel{\text{ind.}}{\sim} N(0, \sigma_\varepsilon^2).$$

Derive an expression for

$$\text{Cov}(y_{ij}, y_{ij'}), \quad j \neq j',$$

and simplify as much as possible. This corresponds to the covariance between repeated measurements on the same subject.

3. Nanette Corrigan is an obstetrician working with and researching the inhabitants of Mussau Island in Papua New Guinea. A current research project involves risk factors for rubella in new-born babies. A survey involving a random sample of 107 births leads to a logistic regression model of the form

$$\text{logit}\{P(y_i = 1)\} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$$

where, for birth i ,

$$\begin{aligned} y_i &= \text{indicator of rubella,} \\ x_{1i} &= \text{maternal age in years,} \\ \text{and } x_{2i} &= \text{glycemic control index} \end{aligned}$$

for $1 \leq i \leq 107$.

Nanette fits the model using R and obtains the following output:

Call:

```
glm(formula = y ~ x1 + x2, family = "binomial")
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.2029	-0.5420	0.1407	0.5279	2.1464

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-7.94435	1.81804	-4.370	1.24e-05	***
x1	0.16901	0.04726	3.576	0.000348	***
x2	7.27934	1.36447	5.335	9.56e-08	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 148.100 on 106 degrees of freedom
Residual deviance: 84.844 on 104 degrees of freedom
AIC: 90.844

Number of Fisher Scoring iterations: 5

(a) Let

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} \\ \vdots & \vdots & \vdots \\ 1 & x_{1,107} & x_{2,107} \end{bmatrix}$$

and $\hat{\boldsymbol{\beta}}$ be the 3×1 vector of maximum likelihood estimates of β_0 , β_1 and β_2 . Explain how, given \mathbf{X} and $\hat{\boldsymbol{\beta}}$, computation of the standard errors can be performed.

- (b) Nanette decides to perform some diagnostic checks on her model. The first of these involves a normal qq-plot of deviance residuals

$$\mathbf{d} = (d_1, \dots, d_{107})$$

which are given by

$$\mathbf{d} = 2\text{sign}(\mathbf{y} - \hat{\boldsymbol{\mu}}) \sqrt{\mathbf{y}^T \ln(\mathbf{y}/\hat{\boldsymbol{\mu}}) + (\mathbf{1} - \mathbf{y})^T \ln((\mathbf{1} - \mathbf{y})/(\mathbf{1} - \hat{\boldsymbol{\mu}}))}$$

where

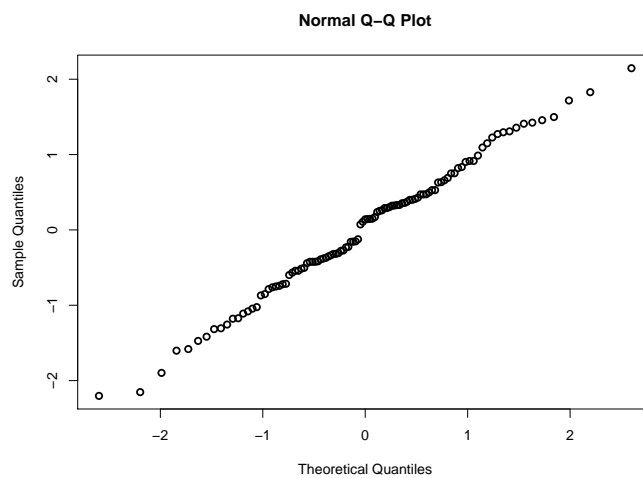
$$\hat{\mu}_i = 1/\{1 + \exp(-\hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})\}$$

The sign function is defined by

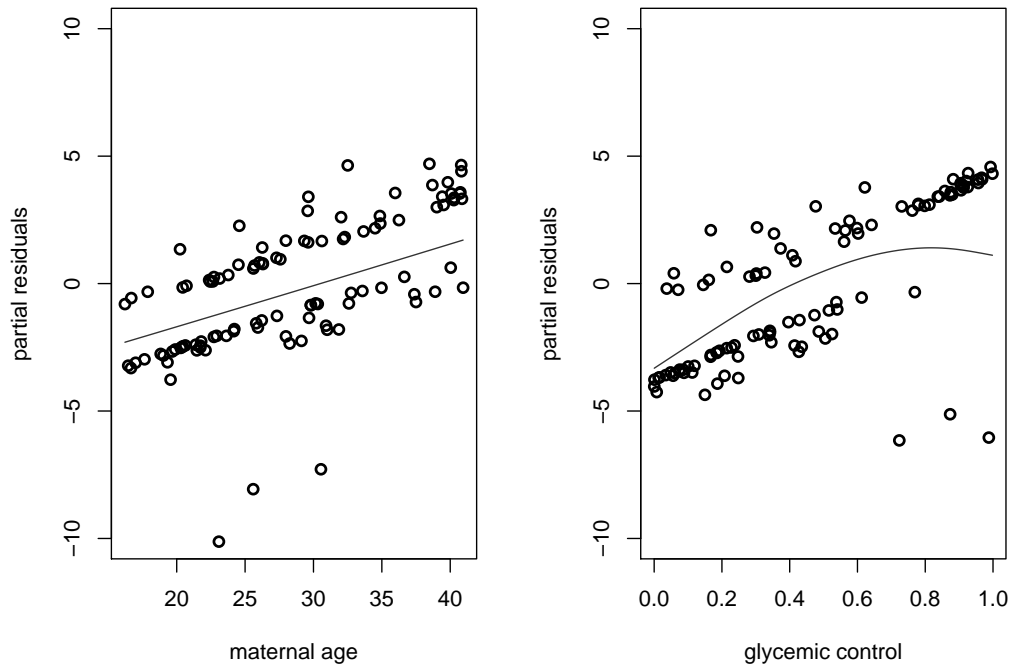
$$\text{sign}(x) = \begin{cases} -1 & \text{if } x < 0 \\ 0 & \text{if } x = 0 \\ 1 & \text{if } x > 0. \end{cases}$$

If $(x_{11}, x_{21}, y_1) = (24, 0.7, 0)$ then what is the value of d_1 ? Simply your answer as much as is possible without the aid of a computer or calculator.

- (c) Nanette obtains the normal qq-plot of the deviance residuals, which is shown below. Briefly comment on this plot in the context of Nanette's analysis.



- (d) Nanette then obtains partial residual plots, which are shown below. Briefly comment on these plots in the context of Nanette's analysis.



(e) The *complementary log-log* regression model for these data is

$$C\{P(y_i = 1)\} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$$

where the function C is given by $C(p) = \ln\{-\ln(1 - p)\}$, $0 < p < 1$. Obtain an expression for the likelihood of $(\beta_0, \beta_1, \beta_2)$ according to this model.

4. (a) Alvin Semba works as a statistician for the *Yundini* wine investment consortium. Yundini recently commissioned a 10-year study where 87 vintages from 3 different countries, Australia, Chile and France, were followed. Quality scores based on expert tasting were recorded every 3 months. For $1 \leq i \leq 87$ and $1 \leq j \leq 40$ let

y_{ij} = quality score for vintage i at the j th recording,

$$c_i = \begin{cases} 1 & \text{if } i\text{th vintage is from Chile} \\ 0 & \text{otherwise,} \end{cases}$$

$$f_i = \begin{cases} 1 & \text{if } i\text{th vintage is from France} \\ 0 & \text{otherwise,} \end{cases}$$

and

a_{ij} = age in years of vintage i at the j th recording.

Alvin decides to fit the additive mixed model

$$y_{ij} = \beta_0 + \beta_c c_i + \beta_f f_i + U_i + \beta_a a_{ij} + \sum_{k=1}^{25} u_k (a_{ij} - \kappa_k)_+ + \varepsilon_{ij} \quad (2)$$

where $\kappa_1, \dots, \kappa_k$ are knots, β_0, β_1 and β_2 are fixed (non-random) regression coefficients, $U_1, \dots, U_{87} \stackrel{\text{ind.}}{\sim} N(0, \sigma_U^2)$ is a random vintage effect intercept and $u_1, \dots, u_{25} \stackrel{\text{ind.}}{\sim} N(0, \sigma_u^2)$ are random spline coefficients and independent of the U_i . Lastly, $\varepsilon_{ij} \stackrel{\text{ind.}}{\sim} N(0, \sigma_\varepsilon^2)$ are random errors independent of the U_i and u_i .

- (i) Alvin would like to fit (2) using mixed model software. This involves writing it in the form:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$$

$$\begin{bmatrix} \mathbf{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} \sim N \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix} \right)$$

What are the matrices \mathbf{X} , $\boldsymbol{\beta}$, \mathbf{Z} , \mathbf{u} , \mathbf{G} and \mathbf{R} for Alvin's model (i.e. (2))?

- (ii) It can be shown that the maximum likelihood estimator for $\boldsymbol{\beta}$ satisfies

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{y}$$

where $\mathbf{V} = \text{Cov}(\mathbf{y})$. Use this to obtain an expression for

$$\text{Cov}(\tilde{\boldsymbol{\beta}})$$

and simplify as much as possible.

- (iii) Describe how Alvin should assess the difference in quality between French and Australian wines. Your answer should include a (point) estimator for the mean difference and a measure of the estimate's variability.

- (b) Let (x_i, y_i) , $1 \leq i \leq n$, be a set of predictor/response pairs. The nonparametric regression model for these data is

$$y_i = f(x_i) + \varepsilon_i$$

where f is some smooth function ε_i . The vector of fitted values for a penalised spline smoother can be written in the form

$$\hat{\mathbf{y}}_\lambda = \mathbf{C}(\mathbf{C}^T \mathbf{C} + \lambda \mathbf{D})^{-1} \mathbf{C}^T \mathbf{y}$$

where \mathbf{C} is an $n \times L$ design matrix containing polynomial and spline basis functions of the x_i s and \mathbf{D} is a $L \times L$ diagonal matrix of zeroes and ones.

The generalized cross-validation criterion for choosing λ is

$$\text{GCV}(\lambda) = \frac{\text{RSS}(\lambda)}{\{n - \text{df}(\lambda)\}^2}$$

where $\text{RSS}(\lambda) = \|\mathbf{y} - \hat{\mathbf{y}}_\lambda\|^2$ is the residual sum of squares and

$$\text{df}(\lambda) = \text{tr}\{\mathbf{C}(\mathbf{C}^T \mathbf{C} + \lambda \mathbf{D})^{-1} \mathbf{C}^T\} \quad (3)$$

- (i) Obtain an explicit expression for $\text{RSS}(\lambda)$ involving quadratic forms in \mathbf{y} (i.e. those of the form $\mathbf{y}^T \mathbf{A} \mathbf{y}$) and simplify it as much as possible.
- (ii) The degrees of freedom expression (3) suffers from the problem that it involves the $n \times n$ matrix $\mathbf{C}(\mathbf{C}^T \mathbf{C} + \lambda \mathbf{D})^{-1} \mathbf{C}^T$. This can lead to storage problems when n is large. Show how matrix algebraic manipulations can overcome this problem.

5. Consider the (frequentist) Poisson mixed model

$$[\mathbf{y}_i|\mathbf{u}] \stackrel{\text{ind.}}{\sim} \text{Poisson}\{\exp(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u})_i\},$$
$$[\mathbf{u}] \sim N(\mathbf{0}, \sigma^2\mathbf{I}).$$

where \mathbf{u} is a $q \times 1$ random effects vector. The likelihood for estimation of $(\boldsymbol{\beta}, \sigma^2)$ is

$$\mathcal{L}(\boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-q/2} J(\boldsymbol{\beta}, \sigma^2) \prod_{i=1}^n (1/y_i!)$$

where

$$J(\boldsymbol{\beta}, \sigma^2) = \int_{\mathbb{R}^q} \exp \left\{ \mathbf{y}^T(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) - \mathbf{1}^T e^{\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}} - \frac{\mathbf{u}^T \mathbf{u}}{2\sigma^2} \right\} d\mathbf{u}.$$

Use Laplace's Method to approximate the log-likelihood $\ell(\boldsymbol{\beta}, \sigma^2) = \ln\{\mathcal{L}(\boldsymbol{\beta}, \sigma^2)\}$. All working should be shown.

6. (a) Consider the Bayesian normal linear regression model

$$[\mathbf{y}|\boldsymbol{\beta}, \sigma^2] \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$$

with prior distributions

$$[\boldsymbol{\beta}] \sim N(\mathbf{0}, \sigma_\beta^2\mathbf{I}), \quad [\sigma^2] \sim \text{IG}(A, B).$$

- (i) Derive the full conditional distributions

$$[\boldsymbol{\beta}|\mathbf{y}, \sigma^2] \quad \text{and} \quad [\sigma^2|\boldsymbol{\beta}, \mathbf{y}]$$

- (ii) Let $\boldsymbol{\beta}^{[1]}$ and $(\sigma^2)^{[1]}$ be initial values of $\boldsymbol{\beta}$ and σ^2 respectively, assumed to be drawn from the posterior distributions $[\boldsymbol{\beta}|\mathbf{y}]$ and $[\sigma^2|\mathbf{y}]$. Write down the set of steps required to obtain the Gibbs samples

$$\boldsymbol{\beta}^{[2]}, (\sigma^2)^{[2]}, \boldsymbol{\beta}^{[3]} \quad \text{and} \quad (\sigma^2)^{[3]}.$$

(If you do not have a solution to (i) then write your answer in general terms).

- (b) Consider the Bayesian *probit* regression model: independently for $1 \leq i \leq n$,

$$P(y_i = 1|\boldsymbol{\beta}) = \Phi((\mathbf{X}\boldsymbol{\beta})_i), \quad [\boldsymbol{\beta}] \sim N(0, 10^8\mathbf{I}),$$

where $y_i \in \{0, 1\}$ is binary, $(\mathbf{X}\boldsymbol{\beta})_i$ is the i th entry of $\mathbf{X}\boldsymbol{\beta}$, \mathbf{X} is a design matrix (suitably standardised) and $\boldsymbol{\beta}$ is a vector of regression coefficients. Note that Φ is the cumulative distribution function of the standard normal distribution. The posterior $[\boldsymbol{\beta}|\mathbf{y}]$ does not have an explicit form, so a common step towards Bayesian inference is to introduce an $n \times 1$ *auxiliary* random vector $\boldsymbol{\alpha} = [\alpha_1, \dots, \alpha_n]^T$ and re-write the model as:

$$\begin{aligned} [\mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\beta}] &= \prod_{i=1}^n I(\alpha_i \geq 0)^{y_i} I(\alpha_i < 0)^{1-y_i}, \\ [\boldsymbol{\alpha}|\boldsymbol{\beta}] &\sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{I}), \\ [\boldsymbol{\beta}] &\sim N(0, 10^8\mathbf{I}). \end{aligned}$$

where, for a logical condition \mathcal{P} ,

$$I(\mathcal{P}) = \begin{cases} 1 & \text{if } \mathcal{P} \text{ is true} \\ 0 & \text{if } \mathcal{P} \text{ is false.} \end{cases}$$

- (i) Determine the full conditional distribution $[\boldsymbol{\beta}|\boldsymbol{\alpha}, \mathbf{y}]$.
(ii) Obtain an expression for full conditional distribution $[\boldsymbol{\alpha}|\boldsymbol{\beta}, \mathbf{y}]$. The normalising constant can be ignored.
(iii) If $y_1 = 1$ then what is the distribution of $[\alpha_1|\boldsymbol{\beta}, \mathbf{y}]$?

POSSIBLY USEFUL FACTS

The $X \sim \text{IG}(A, B)$ then X has density function

$$[x; A, B] = \frac{B^A}{\Gamma(A)} x^{-A-1} e^{-B/x}, \quad x > 0.$$

If the $d \times 1$ random vector \mathbf{X} has a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ then the density function of \mathbf{x} is

$$\phi_{\boldsymbol{\Sigma}}(\mathbf{x} - \mathbf{u}) = (2\pi)^{-d/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\}.$$

For vectors \mathbf{x} and \mathbf{a} and matrix \mathbf{A} ,

$$\phi_{\boldsymbol{\Sigma}}(\mathbf{x} - \mathbf{A}\mathbf{a})\phi_{\boldsymbol{\Lambda}}(\mathbf{a}) = c\phi_{(\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{A} + \boldsymbol{\Lambda}^{-1})^{-1}}(\mathbf{a} - (\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{A} + \boldsymbol{\Lambda}^{-1})^{-1}\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{x})$$

where the factor c depends only on $\mathbf{x}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}$ and \mathbf{A} , but not on \mathbf{a} .

If \mathbf{x} is a random vector, \mathbf{A} is a constant matrix and \mathbf{c} is a constant vector whose dimensions are such that $\mathbf{A}\mathbf{x} + \mathbf{c}$ is defined then

$$E(\mathbf{A}\mathbf{x} + \mathbf{c}) = \mathbf{A}E(\mathbf{x}) + \mathbf{c} \quad \text{and} \quad \text{Cov}(\mathbf{A}\mathbf{x} + \mathbf{c}) = \mathbf{A}\text{Cov}(\mathbf{x})\mathbf{A}^T.$$

The updating step for the Newton-Raphson Method for maximising the smooth function $S(\mathbf{x})$ over $\mathbf{x} \in \mathbb{R}^d$ is:

$$\mathbf{x}_{i+1} = \mathbf{x}_i - \text{HS}(\mathbf{x}_i)^{-1}\{\text{DS}(\mathbf{x}_i)\}^T.$$

For square matrices \mathbf{A} and \mathbf{B} , $|\mathbf{A}\mathbf{B}| = |\mathbf{A}||\mathbf{B}|$.

If \mathbf{A} is a $d \times d$ matrix and $a \in \mathbb{R}$ then $|a\mathbf{A}| = a^d|\mathbf{A}|$.

If

$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}\right)$$

is a general partitioned normal random vector, then the marginal distribution of \mathbf{x}_2 is $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$ and the conditional distribution of \mathbf{x}_2 given \mathbf{x}_1 is

$$[\mathbf{x}_2|\mathbf{x}_1] \sim N(\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1), \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}).$$

Multivariate Laplace's Method

$$\int_{\mathbb{R}^d} e^{h(\mathbf{x})} d\mathbf{x} \simeq e^{h(\mathbf{x}_0)} \sqrt{\frac{(2\pi)^d}{|-\text{H}h(\mathbf{x}_0)|}}$$

where \mathbf{x}_0 is the solution to $\text{D}h(\mathbf{x}) = \mathbf{0}$.