First name:

Last name:

Student ID number:

Statistical Modelling
Exam 23/01/2025

Instructions for the exam:
- Notation: bold symbols indicate vectors,
- The Gaussian distribution is denoted as $\mathrm{N}(\mu, \sigma^2)$, with μ mean parameter and σ^2 variance.
- When performing statistical tests, explicitly write: the system of hypotheses, test statistic and its distribution, observed value, reject region and conclusion.

Exercise 1

On $n = 20$ statistical units we observe the values of two continuous numeric variables (y_i, x_i), $i = 1, \dots, n$. To these data, it is fitted the linear regression model

$$Y_i = \beta_1 + \beta_2(x_i - \bar{x}) + \beta_3(x_i - \bar{x})^2 + \varepsilon_i$$

with $\bar{x} = (1/20)\sum_{i=1}^{20} x_i$, and $(\varepsilon_1, \dots, \varepsilon_{20})$ independent random variables with Gaussian distribution $\mathrm{N}(0, 4)$.

Answer the following questions:

a) Write the parameter and sample space.

b) Express the model in matrix form: $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, explicitly stating how $\boldsymbol{Y}$, X, $\boldsymbol{\beta}$, and $\boldsymbol{\varepsilon}$ are defined and their dimensions. Write the distribution of $\boldsymbol{Y}$ and $\boldsymbol{\varepsilon}$.

c) Write the likelihood and log-likelihood for the parameters of the model.

d) Knowing that,

$$(X^T X)^{-1} = \begin{bmatrix} 0.8 & -1.9 & 2.5 \\ -1.9 & 5.9 & -9.4 \\ 2.5 & -9.4 & 18.7 \end{bmatrix}, \qquad X^T\boldsymbol{y} = \begin{bmatrix} 21 \\ 14 \\ 4 \end{bmatrix}, \qquad \boldsymbol{y}^T\boldsymbol{y} = 473.78,$$

and that the sample means of $\boldsymbol{y}$ and $\boldsymbol{x}$ are, respectively, $\bar{y} = 0.2$ and $\bar{x} = 8$, obtain the maximum likelihood estimates of the regression parameters.

e) Write the exact distribution of the estimator $\hat{B}_2$ of β_2.

f) Perform a test to evaluate whether it is reasonable to keep the quadratic term.

g) Write the definition of the coefficient of determination R^2. The R^2 of the fitted model is equal to 0.122, how do you interpret this value?

h) Perform a test about the overall significance of the model using a 10% significance level.

i) Let $\boldsymbol{e} = \boldsymbol{y} - X^T\hat{\boldsymbol{\beta}}$ be the vector of the residuals. Indicate which of the following identities are true and motivate the answer:

$$\sum_{i=1}^{20} e_i = 0, \qquad \sum_{i=1}^{20} e_i x_i = 0, \qquad \sum_{i=1}^{20} e_i x_i = \bar{x}\bar{e}, \qquad \sum_{i=1}^{20} e_i(x_i - \bar{x})^2 = 0.$$

Exercise 2

The *Pima* dataset was collected by the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements. In particular, the $n = 724$ patients in this dataset are females at least 21 years old of Pima heritage.

The datasets has one response variable (`diabetes`: 1 = positive; 0 = negative), and it is known that, of these women, 249 have diabetes, while 475 do not.
Moreover, we have the following medical predictor variables:

- `pregnant`: Presence of present/past pregnancies: 0 = no pregnancies; 1 = at least one pregnancy.
- `glucose` : Plasma glucose concentration, numeric.
- `pressure`: Diastolic blood pressure (mm Hg), numeric.
- `BMI` : Body mass index, numeric.
- `age` : Age (years), numeric.

Fitting a logistic regression on R returns the following output ("model A"):

	Estimate	Std.Error	z value	Pr(>|z|)
(Intercept)	-8.9267	0.8537	-10.46	0.0000
pregnant	0.2465	0.2931	0.84	0.4004
glucose	0.0349	0.0035	9.92	0.0000
pressure	-0.0078	0.0084	-0.93	0.3515
BMI	0.0941	0.0154	6.09	0.0000
age	0.0328	0.0086	3.81	0.0001

Null deviance: 931.94 on 723 degrees of freedom
Residual deviance: 694.45 on 718 degrees of freedom

Answer the following:

a) Write the corresponding theoretical model.

b) Write the likelihood and log-likelihood functions for the regression parameters of the model.

c) Provide the interpretation of the `age` and `pregnant` coefficients.

d) Is it reasonable to remove the `pregnant` variable from the regression? Why?

e) Define the concept of "odds" and how to interpret it.

A new model ("model B") is then fitted removing the `pregnant` and `pressure` variables. This model returns the following output:

	Estimate	Std.Error	z value	Pr(>|z|)
(Intercept)	-9.0085	0.7261	-12.41	0.0000
glucose	0.0346	0.0035	9.90	0.0000
BMI	0.0884	0.0147	6.01	0.0000
age	0.0317	0.0079	3.99	0.0001

Null deviance: 931.94 on 723 degrees of freedom
Residual deviance: 696.15 on 720 degrees of freedom

f) Perform a test to compare model A and model B using a 5% significance level. Which one do you prefer?

g) According to model B, what is the probability of developing diabetes for a woman aged 25, with a glucose level equal to 99.75 and a BMI of 22?

h) Define the null model. Obtain the estimate of the regression coefficients in this model.

	p						
	0.90	0.95	0.975	0.99	0.995	0.9975	0.999
z_p	1.2816	1.6449	1.9600	2.3263	2.5758	2.8070	3.0902

Table 1: Some quantiles of the Gaussian distribution: $p = \mathbb{P}(Z \leq z_p)$. Columns correspond to probabilities p.

	p						
	0.9	0.95	0.975	0.99	0.995	0.9975	0.999
$t_{2;p}$	1.8856	2.92	4.3027	6.9646	9.9248	14.089	22.3271
$t_{3;p}$	1.6377	2.3534	3.1824	4.5407	5.8409	7.4533	10.2145
$t_{17;p}$	1.3334	1.7396	2.1098	2.5669	2.8982	3.2224	3.6458
$t_{18;p}$	1.3304	1.7341	2.1009	2.5524	2.8784	3.1966	3.6105
$t_{19;p}$	1.3277	1.7291	2.093	2.5395	2.8609	3.1737	3.5794
$t_{20;p}$	1.3253	1.7247	2.086	2.528	2.8453	3.1534	3.5518

Table 2: Some quantiles of the t distribution: $p = \mathbb{P}(T \leq t_{\alpha;p})$ with $T \sim t_\alpha$. Columns correspond to probabilities p. Rows correspond to different degrees of freedom α.

	p						
	0.9000	0.9500	0.9750	0.9900	0.9950	0.9975	0.9990
$f_{1,18;p}$	3.0070	4.4139	5.9781	8.2854	10.2181	12.3208	15.3793
$f_{2,18;p}$	2.6239	3.5546	4.5597	6.0129	7.2148	8.5130	10.3899
$f_{3,18;p}$	2.4160	3.1599	3.9539	5.0919	6.0278	7.0351	8.4875
$f_{1,17;p}$	3.0262	4.4513	6.0420	8.3997	10.3842	12.5525	15.7222
$f_{2,17;p}$	2.6446	3.5915	4.6189	6.1121	7.3536	8.7006	10.6584
$f_{3,17;p}$	2.4374	3.1968	4.0112	5.1850	6.1556	7.2053	8.7269
$f_{1,16;p}$	3.0481	4.4940	6.1151	8.5310	10.5755	12.8201	16.1202
$f_{2,16;p}$	2.6682	3.6337	4.6867	6.2262	7.5138	8.9179	10.9710
$f_{3,16;p}$	2.4618	3.2389	4.0768	5.2922	6.3034	7.4027	9.0059

Table 3: Some quantiles of the F distribution: $p = \mathbb{P}(F \leq f_{\alpha,\beta;p})$ with $F \sim F_{\alpha,\beta}$. Columns correspond to probabilities p. Rows correspond to different degrees of freedom α and β.

	p						
	0.9	0.95	0.975	0.99	0.995	0.9975	0.999
$\chi^2_{2;p}$	4.6052	5.9915	7.3778	9.2103	10.5966	11.9829	13.8155
$\chi^2_{4;p}$	7.7794	9.4877	11.1433	13.2767	14.8603	16.4239	18.4668
$\chi^2_{6;p}$	10.6446	12.5916	14.4494	16.8119	18.5476	20.2494	22.4577

Table 4: Some quantiles of the χ^2 distribution: $p = \mathbb{P}(\chi^2 \leq \chi^2_{\alpha;p})$ with $\chi^2 \sim \chi^2_\alpha$. Columns correspond to probabilities p. Rows correspond to different degrees of freedom α.