Statistical Modelling
Exam 20/02/2025

Exercise 1

The `chickwts` dataset collects data from an experiment conducted to measure and compare the effect of various feed supplements on the growth rate of chickens.
Specifically, the weight in grams of several chickens (the y variable) was measured. The chickens were fed 5 different supplements (`feed` variable):

(1) casein
(2) horse bean,
(3) linseed,
(4) soybean,
(5) sunflower.

Each type of feed was given to 10 chickens. Moreover, it is known that the sample mean and sample variances of the chickens' weight, for each feed, are:

(1) $\bar{y}_1 = 326.8$ $(s_1^2 = 4871.9)$,
(2) $\bar{y}_2 = 160.2$ $(s_2^2 = 1491.9)$,
(3) $\bar{y}_3 = 211.0$ $(s_3^2 = 2894.0)$,
(4) $\bar{y}_4 = 267.4$ $(s_4^2 = 1992.3)$.
(5) $\bar{y}_5 = 333.2$ $(s_5^2 = 2768.2)$.

Finally, the sample variance of y is $s^2 = 7112.7$.

We fit a Gaussian linear regression model to these data (referred to as "Model A"). The residual sum of squares of this model is SSE = 126165.2

Answer the following questions:

1. Assume that the observations are sorted according to the feed type, and that dummy variables are used to encode them, with `casein` being the reference category.
   Write the statistical model corresponding to the analysis (assuming that the model includes the intercept). 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 $\boldsymbol{\varepsilon}$.
2. Compute the matrix $X^T X$.
3. Write the parameter and sample space.
4. Define and explain the sum of squares decomposition, compute the missing quantities.
5. Perform a statistical test for the hypothesis that the type of feed does not have an effect on the chickens' weight using a 5% significance level.
6. Obtain the maximum likelihood estimate of $\boldsymbol{\beta}$.

7. Knowing that

$$(X^TX)^{-1} = \begin{bmatrix} 0.10 & -0.10 & -0.10 & -0.10 & -0.10 \\ -0.10 & 0.20 & 0.10 & 0.10 & 0.10 \\ -0.10 & 0.10 & 0.20 & 0.10 & 0.10 \\ -0.10 & 0.10 & 0.10 & 0.20 & 0.10 \\ -0.10 & 0.10 & 0.10 & 0.10 & 0.20 \end{bmatrix}$$

write the distribution of the maximum likelihood estimator $\hat{\boldsymbol{B}}$ and the marginal distribution of $\hat{B}_{horsebean}$.

8. Knowing that the estimate $\hat{\beta}_{horsebean} = -166.60$, perform a statistical test about the significance of $\beta_{horsebean}$ using a 5% significance level. What is the meaning of this test in the context of the study about the feed type?

9. Compute the coefficient of determination R^2 of this model and explain it.

10. Let $\boldsymbol{e} = \boldsymbol{y} - X\hat{\boldsymbol{\beta}}$ be the vector of the residuals. State which of the following identities are satisfied and motivate the answer:

$$(a)\ \sum_{i=1}^{20} e_i = 0 \qquad (b)\ \sum_{i=1}^{5} e_i = 0 \qquad (c)\ \sum_{i=1}^{10} e_i = 0 \qquad (d)\ \sum_{i=11}^{40} e_i = 0$$

An additional variable, `vit`, is introduced, indicating the amount of a vitamin supplement in micrograms. Denote this model with "Model B". Fitting this model yields the following output:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	287.2494	27.1730	10.57	0.0000
horsebean	-161.0150	23.2958	-6.91	0.0000
linseed	-120.1634	23.2170	-5.18	0.0000
soybean	-61.7686	23.1296	-2.67	0.0106
sunflower	-4.1520	23.8089	-0.17	0.8624
vit	7.6204	4.1847	1.82	0.0754
SSE = 120260.2				

11. Perform a statistical test to compare Models A and B using a 5% significance level.

12. Obtain the coefficient R^2_B of model B. Is it sufficient to compare the R^2 values of the two models instead of performing the formal test? Why or why not?

13. Provide the interpretation of the `linseed` and `vit` coefficients.

14. Compute the estimated regression lines for one chicken given `casein` and another given `linseed`. Sketch both lines.

Exercise 2

The `UCBAdmissions` dataset contains data on the admission status of 4,526 applicants to graduate programs at Berkeley across six departments in 1973. The variables included in the dataset are:

- `admitted`: admission status: 1 = admitted; 0 = rejected.
- `gender`: categorical variable with levels `male` and `female`.
- `dep`: categorical variable indicating the department applied to, with levels `A`, `B`, `C`, `D`, `E`, `F`.

We fit a logistic regression model to investigate whether gender and department affect the probability of admission. The output from R is as follows:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.6819	0.0991	6.88	0.0000
dep B	-0.0434	0.1098	-0.40	0.6928
dep C	-1.2626	0.1066	-11.84	0.0000
dep D	-1.2946	0.1058	-12.23	0.0000
dep E	-1.7393	0.1261	-13.79	0.0000
dep F	-3.3065	0.1700	-19.45	0.0000
gender male	-0.0999	0.0808	-1.24	0.2167

Null deviance: 6044.3 on 4525 degrees of freedom
Residual deviance: 5187.5 on 4519 degrees of freedom

Answer the following questions:

1. Write the model formulation and state its assumptions.
2. Write the likelihood and log-likelihood functions for the model parameters.
3. Provide the interpretation of the `dep D` coefficient.
4. Which variables appear not to significantly affect the probability of admission? Justify your answer.
5. Perform a test of overall significance at the 10% significance level.