

Problem Sheet 7

MATH50011
Statistical Modelling 1

Week 8

Lecture 13: Properties of Least Squares

1. Consider an “error in the variables” model in which there is a true relationship between Y and w given by $Y_i = \beta_1 + \beta_2 w_i + \epsilon_i$ with $E(\epsilon_i) = 0$ and $\text{Var}(\epsilon_i) = \sigma^2$ independent for $i = 1, \dots, n$. Suppose that rather than observe w_i , we have X_i , an imprecise measurement given by $X_i = \alpha_1 + \alpha_2 w_i + \delta_i$ with $E(\delta_i) = 0$ and $\text{Var}(\delta_i) = \tau^2$ which are independent and independent of the ϵ_i s. The parameters α_1 and α_2 are unknown.

We fit a regression model based on $E(Y_i|X_i) = \gamma_1 + \gamma_2 X_i$. Show that the least squares estimator $\hat{\gamma}$ is biased for estimating β , even when $\alpha_0 = 0$ and $\alpha = 1$.

2. Consider a linear model with a p -dimensional parameter vector β . For a deterministic vector $c \in \mathbb{R}^p$ we know that $c^T \hat{\beta}$, where $\hat{\beta}$ is the least squares estimator, is a linear unbiased estimator for $c^T \beta$.

In linear models of your choice and for vectors c of your choice, give examples for other unbiased linear estimators for $c^T \beta$. Quantify the loss in precision (measured by increase in MSE) of some of those estimators (what assumptions do you need to make for this?).

(Hint: if you are not sure where to start, consider the Week 6 unseen problems.)

3. (a) Compute the projection matrix onto $\text{span}((1, 1, 1, 1)^T, (0, 0, 1, 1)^T)$ in \mathbb{R}^4 .
(b) Compute the projection matrix onto $\text{span}((1, 0, 0)^T, (1, 1, 1)^T, (0, 0, 2)^T)$ in \mathbb{R}^3 .
(c) Compute the projection matrix onto $\text{span}((1, \dots, 1)^T)$ in \mathbb{R}^n .

(d) Compute the projection matrix onto $\text{span}((0, \dots, 0)^T)$ in \mathbb{R}^n .

What is the rank of these matrices? What are their eigenvalues (including their multiplicities)?

Lecture 14: Fitted Values, Residuals

4. Suppose we are interested in the relationship between the height Y of tomato plants one month after being potted in soil a , b , or c .

Plants $1, \dots, n$ are potted in soil a . Plants $n + 1, \dots, 2n$ are potted in soil b . Plants $2n + 1, \dots, 3n$ are potted in soil c .

(a) Write a linear model for this setting of the form

$$E(Y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

where x_{i1} and x_{i2} are binary variables taking on the values 0 and 1.

(b) Express the parameter vector for your model in terms of the mean heights for a tomato plant grown in soil a and/or b and/or c .

(c) Find the least squares estimate of β for your model.

(d) Express the fitted values for your model in terms of appropriate sample means based on the Y_i .

(Harder: how many solutions are there to (a)? How do the fitted values change in each case?)

5. Consider a simple linear regression model, $EY_i = \beta_1 + \beta_2 x_i$ ($i = 1, \dots, n$), where β_1 and β_2 are unknown, the second order assumptions hold and $\text{Var}(Y_i) = \sigma^2 > 0$. Suppose moreover that at least two of the x_i are distinct. Let $e = (e_1, \dots, e_n)^T$ be the vector of residuals. Compute $\text{Cov}(e)$.

6. In a linear model satisfying the second order assumptions, $E(Y) = \beta_1 x_1 + \dots + \beta_p x_p$, where x_1, \dots, x_p are the p columns of the design matrix. For each of the two statements below, state whether it is true or false, justifying your answer in each case.

(a) If the vectors x_1, \dots, x_p are mutually orthogonal, then the residuals are uncorrelated.

(b) If a is a constant vector which is orthogonal to each x_i , then a is orthogonal to the vector of residuals.

7. In a study of the effect of thermal pollution on fish, the proportion of a certain variety of sunfish surviving a fixed level of thermal pollution was determined by Matis and Wehrly (1979) for various exposure times. The following paired data were reported on scaled time (x) versus proportion surviving (y).

x	0.10	0.15	0.2	0.25	0.30	0.35	0.40	0.45	0.5	0.55
y	1.00	0.95	0.95	0.9	0.85	0.7	0.65	0.60	0.55	0.40

(a) Plot the paired data as points in an x - y coordinate system.

(b) Assuming a straight line regression compute the least squares estimates of β_0 (the intercept) and β_1 (the slope).

(c) Estimate $E(Y) = \beta_0 + \beta_1 x$ if the exposure time is $x = 0.325$ units.

(d) Compute the residual sum of squares and give an unbiased estimate of $\sigma^2 = \text{Var}(Y)$.