Link lecture - Lagrange Multipliers

Lagrange multipliers

- provide a method for finding a stationary point of a function, say $f(x, y)$
- when the variables are subject to constraints, say of the form $g(x, y) = 0$
- can need extra arguments to check if maximum or minimum or neither

Links to:

- Calculus unit– the method uses simple properties of partial derivatives
- Statistics unit – can be used to calculate or derive properties of estimators

Lecture will introduce the idea and cover two substantial examples:

- Gauss-Markov Theorem – links to to Statistics 1, §4. Linear regression
- Constrained mles – links to Statistics 1, §3. Maximum likelihood estimation

1. Lagrange multipliers – simplest case

Consider a function f of just two variables x and y. Say we want to find a stationary point of $f(x, y)$ subject to a <u>single</u> constraint of the form $g(x, y) = 0$

- Introduce a <u>single</u> new variable λ – we call λ a Lagrange multiplier
- Find all sets of values of (x, y, λ) such that

$$\nabla f = \lambda \nabla g \quad \text{and} \quad g(x, y) = 0 \quad \text{where} \quad \nabla f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \right)$$

  i.e.

$$\frac{\partial f}{\partial x} = \lambda \frac{\partial g}{\partial x} \quad \text{and} \quad \frac{\partial f}{\partial y} = \lambda \frac{\partial g}{\partial y} \quad \text{and} \quad g(x, y) = 0$$

  (number of equations = original number of variables + number of constraints)

- Evaluate $f(x, y)$ at each of these points. We can often identify the largest/smallest value as the maximum/minimum of $f(x, y)$ subject to the constraint, taking account of whether f is unbounded or bounded above/below.

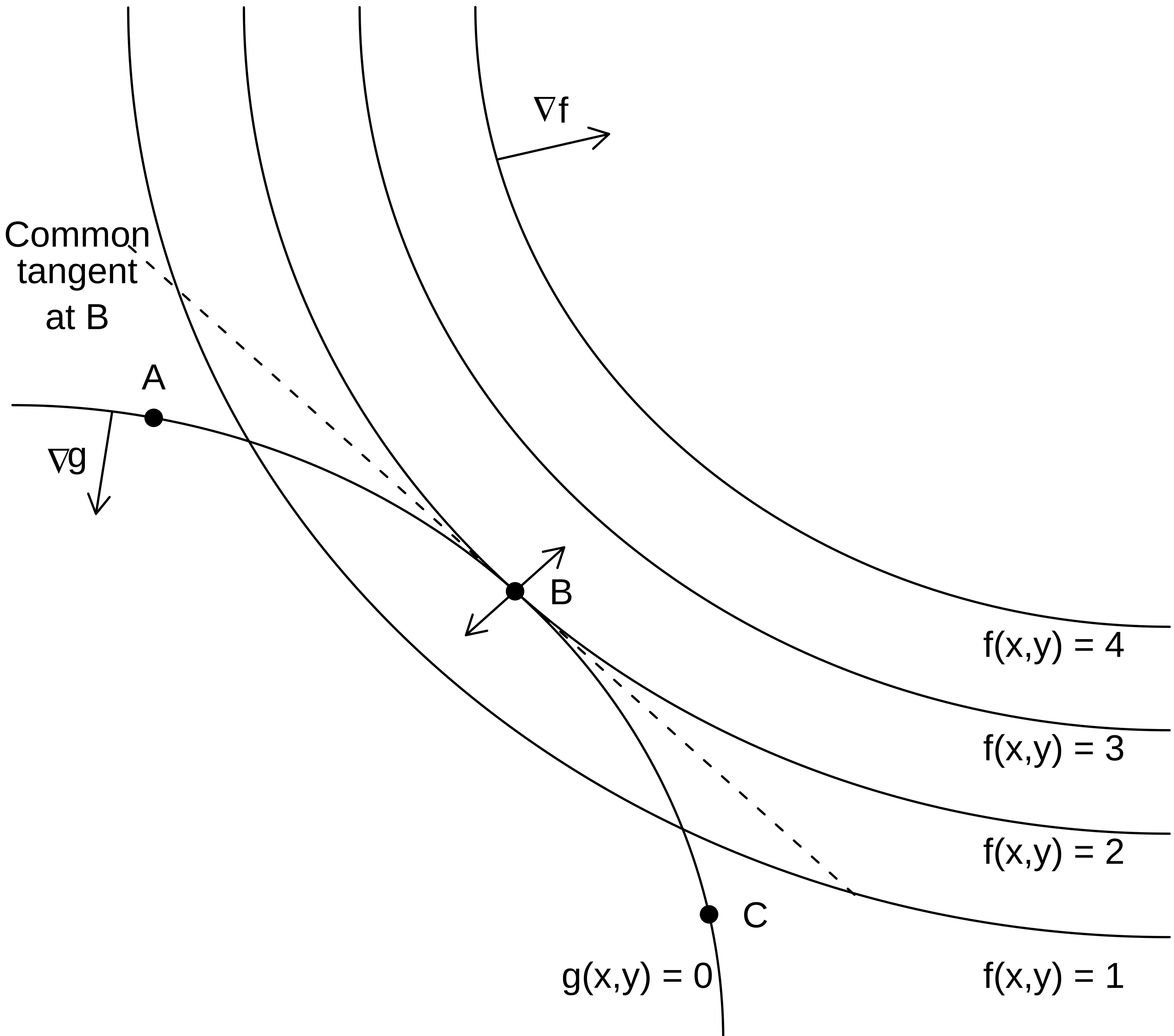
2. Geometric motivation

- Consider finding a local maximum (or local minimum) of $f(x, y)$ subject to a single constraint of the form $g(x, y) = 0$. Recall that a contour of f is a set of points (x, y) for which f takes some given fixed value. Consider how the curve $g(x, y) = 0$ (a contour of g) intersects the contours of f.

- In the following diagram, as we move from A to C along the contour $g(x, y) = 0$, the function $f(x, y)$ first increases then decreases, with a stationary point (here a maximum) at B.

- At the stationary point f and g have a common tangent, so the normal vectors to f and g at that point are parallel, so the gradient vectors are parallel, so

$$\nabla f = \lambda \nabla g \quad \text{for some scalar} \quad \lambda$$


∇f
Common
tangent
at B
A
∇g
B
f(x,y) = 4
f(x,y) = 3
f(x,y) = 2
C
g(x,y) = 0
f(x,y) = 1

3. General number of variables and constraints

The method easily generalises to finding the stationary points of a function f with n variables subject to k independent constraints.

E.g. consider a function $f(x, y, z)$ of <u>three</u> variables x, y, z subject to <u>two</u> constraints $g(x, y, z) = 0$ and $h(x, y, z) = 0$, then:

- at a stationary point ∇f is in the plane determined by ∇g and ∇h
- introduce <u>two</u> Lagrange multipliers, say λ and μ (one per constraint)
- find all sets of values x, y, z, λ, μ satisfying the <u>five</u> (i.e. $3 + 2$) equations

$$\nabla f = \lambda \nabla g + \mu \nabla h \quad \text{and} \quad g(x, y) = 0 \quad \text{and} \quad h(x, y, z) = 0$$

i.e.

$$\frac{\partial f}{\partial x} = \lambda \frac{\partial g}{\partial x} + \mu \frac{\partial h}{\partial x}, \quad \frac{\partial f}{\partial y} = \lambda \frac{\partial g}{\partial y} + \mu \frac{\partial h}{\partial y}, \quad \frac{\partial f}{\partial z} = \lambda \frac{\partial g}{\partial z} + \mu \frac{\partial h}{\partial z}$$

$$g(x, y, z) = 0 \quad \text{and} \quad h(x, y, z) = 0$$

4. Interpretation in terms of the Lagrangian

Again consider the general case of finding a stationary point of a function $f(x_1, \ldots, x_n)$, subject to k constraints $g_1(x_1, \ldots, x_n) = 0, \ldots, g_k(x_1, \ldots, x_n) = 0$

- Introduce k Lagrange multipliers $\lambda_1, \ldots, \lambda_k$
- Define the Lagrangian Λ by

$$\begin{aligned}\Lambda(x, \lambda) &= f(x_1, \ldots, x_n) - \sum_{r=1}^{k} \lambda_r g_r(x_1, \ldots, x_n) \\ &= f(x_1, \ldots, x_n) - \lambda_1 g_1(x_1, \ldots, x_n) - \cdots - \lambda_k g_k(x_1, \ldots, x_n)\end{aligned}$$

- The stationary points of f subject to the constraints $g_1 = 0, \ldots, g_k = 0$ are precisely the sets of values of $(x_1, \ldots, x_n, \lambda_1, \ldots, \lambda_k)$ at which

$$\frac{\partial \Lambda}{\partial x_i} = 0, \quad i = 1, \ldots, n \quad \text{and} \quad \frac{\partial \Lambda}{\partial \lambda_r} = 0, \quad r = 1, \ldots, k$$

  i.e. they are stationary points of the unconstrained function Λ.

5. Example

- maximise $f(x, y) = xy$ subject to $x + y = 1$
  i.e. subject to $g(x, y) = 0$ where $g(x, y) = x + y - 1$
- <u>one</u> constraint so introduce <u>one</u> Lagrange multiplier λ
- compute $\dfrac{\partial f}{\partial x} = y, \quad \dfrac{\partial f}{\partial y} = x, \quad \dfrac{\partial g}{\partial x} = 1, \quad \dfrac{\partial g}{\partial y} = 1$
- and solve the (two + one) equations

$$\frac{\partial f}{\partial x} = \lambda \frac{\partial g}{\partial x} \qquad \text{i.e.} \quad y = \lambda \tag{1}$$

$$\frac{\partial f}{\partial y} = \lambda \frac{\partial g}{\partial y} \qquad \text{i.e.} \quad x = \lambda \tag{2}$$

$$g(x, y) = 0 \qquad \text{i.e.} \quad x + y = 1 \tag{3}$$

- substituting (1) and (2) in (3) gives $2\lambda = 1$, i.e. $\lambda = 1/2$, so from (1) and (2) the function has a stationary point subject to the constraint (here a maximum), at $x = 1/2, y = 1/2$

6. Example

- maximise $f(x, y) = x^2 + 2y^2$ subject to $x^2 + y^2 = 1$
  i.e. subject to $g(x, y) = 0$ where $g(x, y) = x^2 + y^2 - 1$
- one constraint so introduce one Lagrange multiplier λ
- compute $\dfrac{\partial f}{\partial x} = 2x, \quad \dfrac{\partial f}{\partial y} = 4y, \quad \dfrac{\partial g}{\partial x} = 2x, \quad \dfrac{\partial g}{\partial y} = 2y$
- and solve the (two + one) equations

$$\frac{\partial f}{\partial x} = \lambda \frac{\partial g}{\partial x} \qquad \text{i.e.} \quad 2x = \lambda 2x \qquad (1)$$

$$\frac{\partial f}{\partial y} = \lambda \frac{\partial g}{\partial y} \qquad \text{i.e.} \quad 4y = \lambda 2y \qquad (2)$$

$$g(x, y) = 0 \qquad \text{i.e.} \quad x^2 + y^2 = 1 \qquad (3)$$

- (1) $\Rightarrow$ either $\lambda = 1$ or $x = 0$; (2) $\Rightarrow$ either $\lambda = 2$ or $y = 0$
  so possible solutions are $x = 0, \lambda = 2, y = \pm 1$ and $y = 0, \lambda = 1, x = \pm 1$
  where $f(0, \pm 1) = 2$ [max], while $f(\pm 1, 0) = 1$ [min].

7. Gauss-Markov Theorem

A minimum variance property of least squares estimators

- In linear regression the values $y_1, \ldots, y_n$ are assumed to be observed values of random variables $Y_1, \ldots, Y_n$ satisfying the model
  $\mathrm{E}(Y_i) = \alpha + \beta x_i, \quad \mathrm{Var}(Y_i) = \sigma^2, \quad i = 1, \ldots, n$

- A *linear* estimator of β is an estimator of the form $c_1 Y_1 + c_2 Y_2 + \cdots + c_n Y_n$ for some choice of constants $c_1, \ldots, c_n$.

- The *variance* of a linear estimator is $\mathrm{Var}(c_1 Y_1 + \cdots + c_n Y_n) =$
  $c_1^2 \,\mathrm{Var}(Y_1) + \cdots + c_n^2 \,\mathrm{Var}(Y_n) = c_1^2 \sigma^2 + \cdots + c_n^2 \sigma^2 = \sigma^2 \sum_1^n c_i^2$

- A linear estimator is *unbiased* if $\mathrm{E}(\hat{\beta}) = \beta$, i.e.
  $\beta = \mathrm{E}(c_1 Y_1 + c_2 Y_2 + \cdots + c_n Y_n) = c_1 \,\mathrm{E}(Y_1) + \cdots + c_n \,\mathrm{E}(Y_n) =$
  $c_1(\alpha + \beta x_1) + \cdots + c_n(\alpha + \beta x_n) = \alpha(c_1 + \cdots + c_n) + \beta(c_1 x_1 + \cdots + c_n x_n)$
  which requires

$$\sum_1^n c_i = 0 \quad \text{and} \quad \sum_1^n c_i x_i = 1$$

- Thus, for fixed σ^2, a linear estimator that has minimum variance in the class of linear unbiased estimators is obtained by choosing the variables $c_1, \ldots, c_n$ to
  - minimise the objective function $f(c_1, \ldots, c_n) = \sum_1^n c_i^2$
  - subject to the two constraints

$$g(c_1, \ldots, c_n) = \sum_1^n c_i = 0 \quad \text{and} \quad h(c_1, \ldots, c_n) = \sum_1^n c_i x_i - 1 = 0$$

- We introduce two Lagrange multipliers λ and μ
- and compute the $3 \times n$ partial derivatives

$$\frac{\partial f}{\partial c_i} = 2c_i, \qquad \frac{\partial g}{\partial c_i} = 1, \qquad \frac{\partial h}{\partial c_i} = x_i, \qquad\qquad i = 1, \ldots, n$$

- and solve the $(n+2)$ equations

$$\frac{\partial f}{\partial c_i} = \lambda \frac{\partial g}{\partial c_i} + \mu \frac{\partial h}{\partial c_i}, \quad i = 1, \ldots, n, \qquad \sum_1^n c_i = 0, \qquad \sum_1^n c_i x_i - 1 = 0$$

- The first n equations give

$$2c_i = \lambda + \mu x_i, \quad i = 1, \ldots, n$$

- Summing these n equations over $i = 1, \ldots, n$ and using $\sum_1^n c_i = 0$ gives

$$2\sum c_i = n\lambda + \mu \sum x_i \Longrightarrow \lambda = -\mu(\sum x_i)/n = -\mu\bar{x}$$

- On the other hand, multiplying each equation by x_i and then summing over $i = 1, \ldots, n$ gives

$$2\sum c_i x_i = \lambda \sum x_i + \mu \sum x_i^2$$

- Now using $\sum_1^n c_i x_i - 1 = 0$ and using $\lambda = -\mu\bar{x}$ from above gives

$$2 = -\mu\bar{x}\sum x_i + \mu\sum x_i^2 \Longrightarrow \mu = 2/(\sum x_i^2 - n\bar{x}^2)$$

Thus the values of the variables $c_1, \ldots, c_n$ that minimise the variance of the constrained linear estimator $c_1 Y_1 + c_2 Y_2 + \cdots + c_n Y_n$ are the values satisfying the equations

$$c_i = (\lambda + \mu x_i)/2 \quad i = 1, \ldots, n$$

where

$$\lambda = -\mu \bar{x} \quad \text{and} \quad \mu = 2/(\sum x_i^2 - n\bar{x}^2)$$

so

$$c_i = \frac{(x_i - \bar{x})}{\sum x_i^2 - n\bar{x}^2} \quad i = 1, \ldots, n$$

and the resulting estimate is

$$\hat{\beta} = \sum y_i c_i = \sum \frac{y_i(x_i - \bar{x})}{\sum x_i^2 - n\bar{x}^2} = \frac{\sum y_i x_i - n\bar{y}\bar{x}}{\sum x_i^2 - n\bar{x}^2}$$

which is just the least squares estimate.

This gives a simple proof in linear regression case of the Gauss-Markov theorem: The least squares estimator $\hat{\beta}$ has minimum variance in the class of all linear unbiased estimators of β.

8. Maximum likelihood estimates – multinomial distributions

- Consider a statistical experiment in which a sample of size m is drawn from a large population

- assume each observation can take one of four values – say A_1, A_2, A_3 or A_4

- The respective proportions of these values in the population are θ_1, θ_2, θ_3 and θ_4 so $0 < \theta_i < 1,\ j = 1, \ldots, 4$ and $\theta_1 + \theta_2 + \theta_3 + \theta_4 = 1$

- Assume there are m_1 observations with value A_1, m_2 with value A_2, m_3 with value A_3 and m_4 with value A_4 so $m_1 + m_2 + m_3 + m_4 = m$.

- What are the maximum likelihood estimates of $\theta_1, \theta_2, \theta_3$ and θ_4?

  [Note that in this example we use m rather than n to denote the sample size, so as not to clash with the notation in earlier sections where n denoted the number of variables we were optimising over.]

- Here the m_i are observed values of random variables M_i, $i = 1, \ldots, 4$ where the joint distribution of M_1, M_2, M_3 and M_4 is called a *multinomial* distribution

- The joint distribution has probability mass function

$$p(m_1, m_2, m_3, m_4; \theta_1, \theta_2, \theta_3, \theta_4) = \frac{m!}{m_1! m_2! m_3! m_4!} \, \theta_1^{m_1} \, \theta_2^{m_2} \, \theta_3^{m_3} \, \theta_4^{m_4}$$

and so has log likelihood function

$$\ell(\theta_1, \theta_2, \theta_3, \theta_4) = \text{const} + m_1 \log \theta_1 + m_2 \log \theta_2 + m_3 \log \theta_3 + m_4 \log \theta_4$$

where the constant $c = \log m! - (\log m_1! + \log m_2! + \log m_3! + \log m_4!)$

- Since the θ_i are probabilities and must therefore sum to 1, the maximum likelihood estimates $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \hat{\theta}_4$ are the values that maximise the log likelihood $\ell(\theta_1, \theta_2, \theta_3, \theta_4)$, subject to the condition

$$\theta_1 + \theta_2 + \theta_3 + \theta_4 = 1$$

- Thus we want to maximise the objective function

$$\ell(\theta_1, \theta_2, \theta_3, \theta_4) = c + m_1 \log \theta_1 + m_2 \log \theta_2 + m_3 \log \theta_3 + m_4 \log \theta_4$$

- subject to the constraint

$$g(\theta_1, \theta_2, \theta_3, \theta_4) = \theta_1 + \theta_2 + \theta_3 + \theta_4 - 1 = 0$$

- We introduce a single Lagrange multiplier λ
- and compute the partial derivatives

$$\frac{\partial \ell}{\partial \theta_i} = \frac{m_i}{\theta_i}, \qquad \frac{\partial g}{\partial \theta_i} = 1, \qquad i = 1, \ldots, 4$$

- and solve the (four plus one) equations

$$\frac{\partial \ell}{\partial \theta_i} = \lambda \frac{\partial g}{\partial \theta_i} \qquad i = 1, \ldots, 4 \qquad \theta_1 + \theta_2 + \theta_3 + \theta_4 - 1 = 0$$

- The first four equations give

$$\frac{m_i}{\theta_i} = \lambda, \qquad i = 1, \ldots, 4$$

i.e.

$$\theta_1 = \frac{m_1}{\lambda}, \quad \theta_2 = \frac{m_2}{\lambda}, \quad \theta_3 = \frac{m_3}{\lambda}, \quad \theta_4 = \frac{m_4}{\lambda},$$

- Substituting these values into the last equation gives

$$1 = \theta_1 + \theta_2 + \theta_3 + \theta_4 = \frac{m_1}{\lambda} + \frac{m_2}{\lambda} + \frac{m_3}{\lambda} + \frac{m_4}{\lambda} = \frac{(m_1 + m_2 + m_3 + m_4)}{\lambda} = \frac{m}{\lambda}$$

- Putting $\lambda = m$ back into the equations for each θ_i we see that the maximising values (the maximum likelihood estimates) are

$$\boxed{\hat{\theta}_1 = \frac{m_1}{m}, \quad \hat{\theta}_2 = \frac{m_2}{m}, \quad \hat{\theta}_3 = \frac{m_3}{m}, \quad \hat{\theta}_4 = \frac{m_4}{m}}$$