Lecture 26

Multivariate Optimization

This lecture investigates the problem of locating the extrema of a multivariate function $f : X \to \mathbb{R}$ where f is continuously differentiable on $X \subseteq \mathbb{R}^n$. There is no change in the definition of extremum from the univariate case, and the conditions for optimality remain similar in important ways. For example, a first-order necessary condition for an internal extremum is again that the function be locally flat: neither rising nor falling. Now however there a multiple choice variables, and at a stationary point the slope must be zero in every direction.

26.1 Optimality Conditions

Once again the first-order necessary condition identifies stationary points that are potential extrema. And once again, in order to characterize the stationary points, information about the function's curvature is additionally required. This parallels the univariate case. For example, a second order sufficient condition for a stationary point to be a local maximum is once again that the function be locally concave. This lecture assumes that objective functions are twice continuously differentiable and characterizes the local curvature of a function in terms of its second-order partial derivatives.

26.1.1 First-Order Condition

Naturally, if $\hat{\boldsymbol{x}}$ is a maximum of f, changing $\boldsymbol{x}$ cannot increase the value of the function. This implies that all the partial derivatives must be zero at $\hat{\boldsymbol{x}}$: otherwise we could increase the value of f by slightly changing $\boldsymbol{x}$. Recall that the vector of first-order partial derivatives is called the ***gradient*** of f, denoted by Df or even more simply by f'.[1] So a first-order necessary condition for an extremum is that the gradient be zero.

Theorem 26.1 (Necessary Condition for Multivariate Optimization) Consider the behavior of a differentiable function $f : X \to \mathbb{R}$ on an open set $V \subseteq X \subseteq \mathbb{R}_N$. If $\hat{\boldsymbol{x}}$ is a local extremum of f, then $f'[\hat{\boldsymbol{x}}] = \mathbf{0}$.

[1]Other common notations include $f_{\boldsymbol{x}^\top}$, $\partial f/\partial \boldsymbol{x}^\top$, ∇f, or $\nabla_{\boldsymbol{x}} f$.

Proof: For each x_i, consider the univariate mapping $x_i \mapsto f[\hat{x}_1, \ldots, x_i, \ldots, \hat{x}_N]$ and apply the familiar univariate reasoning. This implies $\partial f[\hat{\boldsymbol{x}}]/\partial x_i = 0$ for every x_i. ■

Theorem 26.1 gives us a starting point for finding extrema: impose $\partial f[\boldsymbol{x}]/\partial x_i = 0$ for all i, and find values of the x_i that simultaneously satisfy the entire resulting system of equations. Any point $\boldsymbol{x}$ satisfying this first-order necessary condition is a ***stationary point*** of f.

To explore this necessary condition for an interior extremum, choose an arbitrary vector $\boldsymbol{h} \in \mathbb{R}^N$, and at any $\boldsymbol{x}$, define the univariate function g by

$$g = \alpha \mapsto f[\boldsymbol{x} + \alpha\boldsymbol{h}] - f[\boldsymbol{x}] \tag{26.1}$$

Since f is twice differentiable, the function g will be twice differentiable. Note that

$$g' = \alpha \mapsto f'[\boldsymbol{x} + \alpha\boldsymbol{h}] \cdot \boldsymbol{h} \tag{26.2}$$

Example 26.1 The binary real function $f = \langle x, y\rangle \mapsto 1 - x^2 - y^2$ has the gradient function $f' = \langle x, y\rangle \mapsto \langle -2x, -2y\rangle$. Let $\boldsymbol{h} = \left\langle h_x, h_y\right\rangle$, and for any given $\boldsymbol{x}$ define a *unary* real function $g = \alpha \mapsto f[\boldsymbol{x} + \alpha\boldsymbol{h}] - f[\boldsymbol{x}]$. Then

$$\begin{aligned} g[\alpha] &= \left(1 - (x + \alpha h_x)^2 - (y + \alpha h_y)^2\right) - (1 - x^2 - y^2) \\ &= -(\alpha^2 h_x^2 + \alpha^2 h_y^2 + 2\alpha x h_x + 2\alpha y h_y) \end{aligned}$$

Find the derivative of g:

$$\begin{aligned} g'[\alpha] &= -(2\alpha h_x^2 + 2\alpha h_y^2 + 2x h_x + 2y h_y) \\ &= \left\langle -2(x + \alpha h_x), -2(y + \alpha h_y)\right\rangle \cdot \left\langle h_x, h_y\right\rangle \\ &= f'[\boldsymbol{x} + \alpha\boldsymbol{h}] \cdot \left\langle h_x, h_y\right\rangle \end{aligned}$$

Now suppose $\hat{\boldsymbol{x}}$ maximizes the twice-differentiable function $f : \mathbb{R}^N \to \mathbb{R}$. Define g in terms of *this* value of $\boldsymbol{x}$.

$$g = \alpha \mapsto f[\hat{\boldsymbol{x}} + \alpha\boldsymbol{h}] - f[\hat{\boldsymbol{x}}] \tag{26.3}$$

The function g has a local maximum at $\alpha = 0$. So g must be stationary at 0.

$$g'[0] = \nabla f[\hat{\boldsymbol{x}}] \cdot \boldsymbol{h} = 0 \tag{26.4}$$

Since $\boldsymbol{h}$ is arbitrary, each of the first-order partial derivatives of f must be zero at $\hat{\boldsymbol{x}}$. The same approach works to show the gradient must be zero for an interior minimum.

Consider the differential

$$df_{\boldsymbol{x}} = f'[\boldsymbol{x}] \cdot \mathrm{d}\boldsymbol{x} \tag{26.5}$$

Since the partial derivatives are all zero at an extremum, the differential of f is also zero at $\hat{\boldsymbol{x}}$.

26.1.2 Second-Order Conditions

Continue to explore the possibility that the continuously twice differentiable function $f : \mathbb{X} \to \mathbb{R}$ to has a maximum on an open set $\mathbb{V} \subseteq \mathbb{X} \subseteq \mathbb{R}^N$. Suppose $\hat{\boldsymbol{x}} \in \mathbb{V}$ satisfies the first-order necessary condition. Is it a maximizer, or a minimizer, or neither? Just as in the univariate case, pinning this down requires an exploration of the curvature of the function at $\hat{x}$. If the function is locally concave at $\hat{x}$, it is a local maximum. If it is locally convex at $\hat{x}$, it is a local minimum.

Hessian Matrix

To explore the curvature, let f'' be the matrix of second-order partial derivatives. In terms of the representative $\langle i, j\rangle$-th matrix element,

$$f'' = \boldsymbol{x} \mapsto \left[f_{i,j}[\boldsymbol{x}]\right] \tag{26.6}$$

where $f_{i,j}$ represent the second-order partial derivative with respect to the i-th and j-th arguments.

$$f_{i,j}[\boldsymbol{x}] = \frac{\partial^2 f[\boldsymbol{x}]}{\partial x_i \, \partial x_j} \tag{26.7}$$

This matrix of second-order partials derivatives is the ***Hessian*** matrix. Recall that if f is continuously twice-differentiable then $f_{i,j} = f_{j,i}$, so the Hessian is a symmetric matrix. There are many different notations notations for the Hessian, including the closely related notation

$$f'' = \boldsymbol{x} \mapsto \frac{\partial^2 f[\hat{\boldsymbol{x}}]}{\partial \boldsymbol{x} \partial \boldsymbol{x}^\top} \tag{26.8}$$

which treats the input argument as a column vector. The Hessian of a function f may also be denoted by $\boldsymbol{H}_f$ or $\boldsymbol{D}^2 f$.

Second-Order Condition

At a local maximum the Hessian must be nonpositive definite (negative semidefinite). To explore this, once again consider the univariate function defined in (26.1). Our work on the univariate case tells us that

$$g''[0] = \boldsymbol{h}^\top f''[\hat{\boldsymbol{x}}]\,\boldsymbol{h} \le 0 \tag{26.9}$$

Since $\boldsymbol{h}$ is arbitrary, the Hessian must be nonpositive definite (negative semidefinite). This is a second-order necessary condition for a maximum. The same argument shows that at a minimum the Hessian must be nonnegative definite (positive semidefinite).

The continuing parallels with the univariate case raises a natural question: does a negative definite Hessian at a stationary point imply local concavity of the objective function, and is it a sufficient condition for stationary point to be a strict local maximizer? The answer to both questions is yes, but we will postpone the proof.

Theorem 26.2 Let $\hat{\boldsymbol{x}}$ be a stationary point of the continuously twice-differentiable function $f : X \subseteq \mathbb{R}^N \to \mathbb{R}$. If $f''[\hat{x}]$ is negative definite, then $\hat{\boldsymbol{x}}$ is a strict local maximizer of f. (Similarly, if $f''[\hat{\boldsymbol{x}}]$ is positive definite, then $\hat{\boldsymbol{x}}$ is a strict local minimizer of f.)

Proof: See Simon and Blume (1994, ch.30). ■

Definition 26.1 Let $\boldsymbol{x}$ be a stationary point of the continuously twice-differentiable function $f : X \subseteq \mathbb{R}^N \to \mathbb{R}$. If $\boldsymbol{H}_f[\boldsymbol{x}]$ is indefinite, then $\boldsymbol{x}$ is called a *saddle point* of f.

Theorem 26.3 A saddle point is not a minimizer or a maximizer.

Proof: See Simon and Blume (1994, ch.30). ■

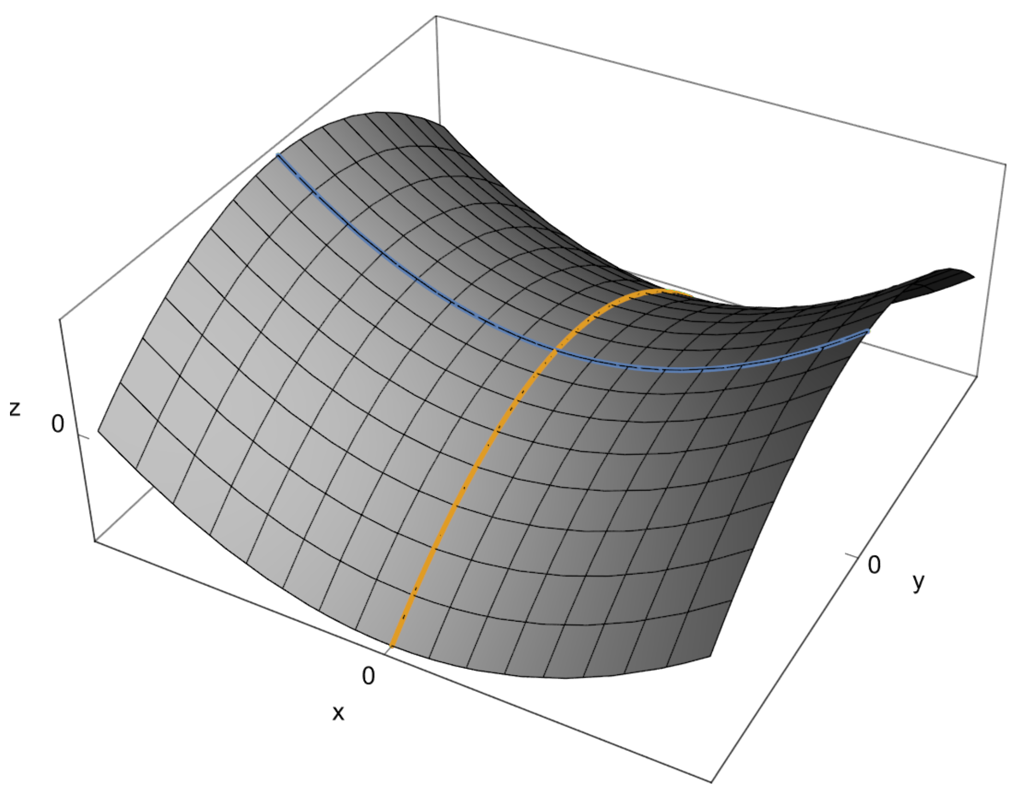

Figure 26.1: Indefinite Quadratic Form

Example 26.2 Consider the quadratic form $f[\boldsymbol{x}] = x_1^2 - x_2^2$. Set $f'[\boldsymbol{x}] = \mathbf{0}$.

$$\langle 2x_1, -2x_2 \rangle = \langle 0, 0 \rangle$$

Solving this equation produces is a stationary point at $\mathbf{0}$. Next, compute the Hessian at $\mathbf{0}$ to find

$$f''[\mathbf{0}] = \begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix}$$

which is indefinite. The stationary point is therefore a saddle point. Specificially, changes in x_1 will increase the value of the function, while changes in x_2 will decrease the value of the function.

Example 26.3 Assume the production function $f : \mathbb{R}^N_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$ is strictly concave in the inputs $\boldsymbol{x}$. Given the prices of output and inputs, a profit maximizing firm wants to

$$\max_{\boldsymbol{x}} p f[\boldsymbol{x}] - \boldsymbol{w}^\top \boldsymbol{x} \tag{26.10}$$

The first-order necessary condition for a maximum is that there be a stationary point, which must satisfy the matrix equation

$$p\, f'[\boldsymbol{x}] - \boldsymbol{w} = \mathbf{0} \tag{26.11}$$

This says that each input is hired until the value of its marginal product equals its factor price. The second-order necessary condition for a maximum is that f'' be nonpositive definite at the stationary point. Note that the Hessian of the objective function is the same as the Hessian of the production function. Since f is globally strictly concave, it is negative definite. This satisfies the second-order necessary condition, and indeed fulfills sufficient conditions for a strict global maximum.

26.2 Application to Curve Fitting

Curve fitting is a core practice in applied social science. A simple approach to curve fitting provides a nice application of the optimization discuss of this lecture. We begin with some data: N observations on a dependent variable y and an explanatory independent variable x. What is the relation between them? Pindyck and Rubinfeld (1997) consider the following approaches to fitting straight lines to the points:

- line from lowest x value to highest x value
- best visual fit
- sum of deviations equals zero (but big positive and negative deviations cancel, plus the criteria does not yield a unique line)

- minimize the sum of the absolute deviations (but is is computationally harder, and perhaps underweights large deviations)

- minimize the sum of the squared deviations (computationally simple, and penalizes large errors heavily)

The last of these is the method of least squares, which is the focus of this section.

26.2.1 Least Squares

The optimization problem involves choosing an intercept a and a slope b for the relationship $y = a + bx$ so that fitted values for the dependent variable lie as close as possible to the actual values. The measure of closeness is the least squares criterion, so the objective function is $f = \langle a, b \rangle \mapsto \sum (y_i - a - bx_i)^2$. We wish to choose a and b to minimize the objective function.

$$\min_{a,b} \sum_{i=1}^{N} (y_i - a - bx_i)^2 \tag{26.12}$$

Notice that the data are already given, and optimization involves choosing only two values (the slope and intercept).

The first-order partial derivatives $f'[a, b]$ are

$$\begin{aligned} \frac{\partial}{\partial a} \sum (y_i - a - bx_i)^2 &= -2 \sum (y_i - a - bx_i) \\ \frac{\partial}{\partial b} \sum (y_i - a - bx_i)^2 &= -2 \sum x_i (y_i - a - bx_i) \end{aligned} \tag{26.13}$$

The first-order necessary condition $f'[a, b] = \mathbf{0}$ is therefore a two-equation system in the variables a and b. Write these as

$$\begin{aligned} \sum_{i=1}^{N} (y_i - a - bx_i) &= 0 \\ \sum_{i=1}^{N} x_i (y_i - a - bx_i) &= 0 \end{aligned} \tag{26.14}$$

These first-order necessary conditions, known as the normal equations, are commonly rewritten as

$$\begin{aligned} aN + b \sum x_i &= \sum y_i \\ a \sum x_i + b \sum x_i^2 &= \sum x_i y_i \end{aligned} \tag{26.15}$$

or better yet as

$$
\begin{aligned}
a + b\bar{x} &= \bar{y} \\
a\bar{x} + b\frac{1}{N}\sum x_i^2 &= \frac{1}{N}\sum x_i y_i
\end{aligned}
\tag{26.16}
$$

The normal equations are two linear equations in the two unknowns a and b. Naturally, they have a simple matrix representation.

$$
\begin{bmatrix} 1 & \bar{x} \\ \bar{x} & \frac{1}{N}\sum x_i^2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \bar{y} \\ \frac{1}{N}\sum x_i y_i \end{bmatrix}
\tag{26.17}
$$

Using the inverse matrix, solve this to get

$$
\begin{aligned}
\begin{bmatrix} a \\ b \end{bmatrix} &= \frac{1}{\frac{1}{N}\sum x_i^2 - \bar{x}^2} \begin{bmatrix} \frac{1}{N}\sum x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \begin{bmatrix} \bar{y} \\ \frac{1}{N}\sum x_i y_i \end{bmatrix} \\
&= \frac{1}{\frac{1}{N}\sum x_i^2 - \bar{x}^2} \begin{bmatrix} \bar{y}\frac{1}{N}\sum x_i^2 - \bar{x}\frac{1}{N}\sum x_i y_i \\ -\bar{x}\bar{y} + \frac{1}{N}\sum x_i y_i \end{bmatrix}
\end{aligned}
\tag{26.18}
$$

After some algebraic manipulation, fine $b = \mathrm{cov}[x, y]/\mathrm{var}[x]$.[2]

What about the second order conditions? Looking at (26.16) it is evident that the Hessian is

$$
N \begin{bmatrix} 1 & \bar{x} \\ \bar{x} & \frac{1}{N}\sum x_i^2 \end{bmatrix}
\tag{26.19}
$$

which is easily shown to be positive definite, ensuring a unique minimum.

Matrix Algebra

Return to the normal equations (26.15). This time, create a vector $\boldsymbol{Y}$ of y_i observations and a matrix $\boldsymbol{X}$ of x_i observations plus a constant. In addition, create a vector $\boldsymbol{\beta}$ of the parameters we are trying to estimate.

$$
\boldsymbol{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \qquad \boldsymbol{X} = \begin{bmatrix} 1 & x_1 \\ & \vdots \\ 1 & x_n \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} a \\ b \end{bmatrix}
\tag{26.20}
$$

[2]

$$
\begin{aligned}
b &= \frac{\frac{1}{N}\sum x_i y_i - \bar{x}\bar{y}}{\frac{1}{N}\sum x_i^2 - \bar{x}^2} = \frac{\sum x_i y_i - N\bar{x}\bar{y}}{\sum x_i^2 - N\bar{x}^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \\
&= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \frac{\frac{1}{N-1}}{\frac{1}{N-1}} = \frac{\mathrm{cov}[x, y]}{\mathrm{var}[x]}
\end{aligned}
$$

The first element of $\boldsymbol{X}$ all ones. Using this notation, write the normal equations as

$$\boldsymbol{X}^{\top}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{0} \tag{26.21}$$

which can be rearranged as

$$(\boldsymbol{X}^{\top}\boldsymbol{X})\boldsymbol{\beta} = \boldsymbol{X}^{\top}\boldsymbol{Y} \tag{26.22}$$

and our solution is

$$\boldsymbol{\beta} = (\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}\boldsymbol{Y} \tag{26.23}$$

26.2.2 Some Properties of the Least Squares Estimator

Consider the implied residuals:

$$\hat{e} = \boldsymbol{Y} - \boldsymbol{X}\hat{\beta} = \boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}\boldsymbol{Y} = [I - \boldsymbol{X}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}]\boldsymbol{Y} \tag{26.24}$$

So

$$\boldsymbol{X}^{\top}\hat{e} = \boldsymbol{X}^{\top}\boldsymbol{Y} - \boldsymbol{X}^{\top}\boldsymbol{X}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}\boldsymbol{Y} = \boldsymbol{X}^{\top}\boldsymbol{Y} - \boldsymbol{X}^{\top}\boldsymbol{Y} = \boldsymbol{0} \tag{26.25}$$

So when the regression includes a constant term, the residuals must sum to zero.

Suppose the true model is

$$\boldsymbol{Y} = \boldsymbol{X}\beta + e \tag{26.26}$$

Consider the expected value of $\hat{\beta}$.

$$\begin{aligned}
\hat{\beta} &= (\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}\boldsymbol{Y} \\
&= (\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}(\boldsymbol{X}\beta + e) \\
&= (\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}\boldsymbol{X}\beta + (\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}e \\
&= \beta + (\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}e
\end{aligned} \tag{26.27}$$

So if $\boldsymbol{X}$ is uncorrelated with e, we have an unbiased estimate of the true coefficient parameter.[3]

26.2.3 Omitted Variables

Suppose we obtain the regression estimates

$$Y = b_0 + b_1 X_1 + U \tag{26.28}$$

[3] We can write the residual vector as

$$\begin{aligned}
\hat{e} &= \boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}\boldsymbol{Y} \\
&= \boldsymbol{X}\beta + e - \boldsymbol{X}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}(\boldsymbol{X}\beta + e) \\
&= \boldsymbol{X}\beta - \boldsymbol{X}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}\boldsymbol{X}\beta + e + \boldsymbol{X}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}e \\
&= [I - \boldsymbol{X}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}]e
\end{aligned}$$

but X_2 should have been in the regression. That is we should have gotten the estimates

$$Y = b_0' + b_1' X_1 + b_2' X_2 + U' \tag{26.29}$$

How are the two estimates related?

We will explore this by considering the supplementary regression of X_2 on X_1:

$$X_2 = c_0 + c_1 X_1 + V \tag{26.30}$$

In every case the regression holds exactly on average. (This follows from the OLS regression, which yields a zero average error.) Therefore the deviations of Y from its mean, y, can be written

$$y = b_1 x_1 + U \tag{26.31}$$

or

$$y = b_1' x_1 + b_2' x_2 + U' \tag{26.32}$$

Note that the errors are unchanged, since they have a zero mean value.

Similarly, the deviation of X_2 from its mean can be written

$$x_2 = c_1 x_1 + V \tag{26.33}$$

Substitute this into (26.32) to get

$$y = b_1' x_1 + b_2'(c_1 x_1 + V) + U' \tag{26.34}$$

Then subtract (26.34) from (26.31) to get

$$\mathbf{0} = (b_1 - b_1' - b_2' c_1) x_1 + U - U' - b_2' V \tag{26.35}$$

The final step is to sum (26.35) over all the observations, recalling the errors U, U', and V are each zero on average.

$$0 = (b_1 - b_1' - b_2' c_1)\bar{x}_1 \tag{26.36}$$

or

$$b_1 = b_1' + b_2' c_1 \tag{26.37}$$

So omission of X_2 from the regression biases the coefficent estimate to an extent that depends on the correlation between the omitted and included variables.

Note: Oksanen (1998) notes that these results hold for any regressions yielding zero average errors.

Note: the derivation works even when b_1' and b_2' are vectors.

26.2.4 Instrumental Variables

Optional reading: Davidson and McKinnon 7.4, Pindyck and Rubinfeld (1997, ch.7)

But what if X is correlated with the error term? Well perhaps we could generate a proxy for X that is not. For example, suppose we have a collection of variables Z that are not correlated with the error term but that we think can proxy X.

$$X = Z\gamma + v \tag{26.38}$$

Then we can produce our X proxy as

$$\hat{X} = Z(Z'Z)^{-1}Z'X = P_Z X \tag{26.39}$$

Now reformulate our original problem as

$$Y = X\beta + u = P_Z X\beta + u + \hat{v}\beta \tag{26.40}$$

and produce the standard least squares estimate of β:

$$\hat{\beta} = (X'P_Z'P_Z X)^{-1}X'P_Z'Y \tag{26.41}$$

Noting that P_Z is symmetric idempotent, we have

$$\hat{\beta} = (X'P_Z X)^{-1}X'P_Z Y \tag{26.42}$$

This is the instrumental variables estimator, where the instruments are Z.

This estimator is also known as the 2SLS estimator, because it can be easily produced as two linear regressions: first regress X on the instruments, then regress Y on the fitted values of X.

26.3 Constrained Optimization

Constrained optimization adds new technical difficulties. Binding constraints generally invalidate the first-order and second-order necessary conditions that we developed in the unconstrained case. This section focuses on equality constraints: in addition to an objective function $f[\boldsymbol{x}]$, there is a constraint that $g[\boldsymbol{x}] = 0$. Both functions are assumed to be continuously differentiable.

26.3.1 First-Order Necessary Conditions

The constrained maximization problem considers only the values that $f[\boldsymbol{x}]$ takes on the constraint set $\{\, \boldsymbol{x} \mid g[\boldsymbol{x}] = 0 \,\}$. If $f[\boldsymbol{x}]$ achieves an extremum (on this set) at the point $\hat{\boldsymbol{x}}$, then the gradients of the functions are collinear at $\hat{\boldsymbol{x}}$. That is, for some number λ,

$$f'[\hat{\boldsymbol{x}}] = \lambda\, g'[\hat{\boldsymbol{x}}] \tag{26.43}$$

Consider a proof for the special case of a binary objective function f and a single equality constraint in the two choice variables.[4] The objective is to pick x and y so as to maximize $f[x, y]$, subject to the constraint that $g[x, y] = 0$.

[4] The proof strategy draws on Edwards (1973, p.93).

Begin by focusing on the constraint. Assume satisfaction of the requirements of implicit function theorem, such that the constraint function implies the implicit function γ satisfying $g\big[x, \gamma[x]\big] \equiv 0$. That is, for each x, $\gamma[x]$ is the value of y that ensure the constraint is satisfied. Therefore chaning x has no effect on the value of $g\big[x, \gamma[x]\big]$. Making use of the chain rule yields

$$0 = g'\big[x, y[x]\big] \cdot \begin{bmatrix} 1 \\ y'[x] \end{bmatrix} \tag{26.44}$$

The vector $\langle 1, y'[x] \rangle$ is the direction of movement that will not lead to a constraint violation.

Example 26.4 Consider a budget constraint for a two-good consumer choice problem: $p_x x + p_y y = w$. Define $g = \langle x, y \rangle \mapsto p_x x + p_y y - w$, which has the particularly simple gradient $g' = \langle x, y \rangle \mapsto \big\langle p_x, p_y \big\rangle$. Solve $g[x, y]$ for the implicit function $\gamma = x \mapsto (w - p_x x)/p_y$. Then $g[x, \gamma[x]] = p_x x + p_y((w - p_x x)/p_y) - w \equiv 0$. Correspondingly, $(\mathrm{d}/\mathrm{d}x) g[x, \gamma[x]] = g_1 \cdot 1 + g_2 \cdot (\gamma'[x]) = \langle g_1, g_2 \rangle \cdot \big\langle 1, -p_x/p_y \big\rangle = \big\langle p_x, p_y \big\rangle \cdot \big\langle 1, -p_x/p_y \big\rangle = 0$.

Next, suppose the binary objective function f attains a constrained maximum at $\langle \hat{x}, \hat{y} \rangle$. Consider the unary function $h = x \mapsto f\big[x, \gamma[x]\big]$, which provides the constrained values of f as x varies (and therefore as $\gamma[x]$ varies). Since the unary function attains a maximum, $h'[\hat{x}] = 0$. That is,

$$f'[\hat{x}, \hat{y}] \cdot \begin{bmatrix} 1 \\ y' \end{bmatrix} = 0 \tag{26.45}$$

Example 26.5 Consider a two-good utility function, $u \colon \mathbb{R}^2_{\geq 0} \to \mathbb{R}$. After imposing the budget constraint as discussed above, produce the unary function $x \mapsto u[x, \gamma[x]]$. At a constrained maximum, the derivative of this unary function is 0. Set $(\mathrm{d}/\mathrm{d}x) u[x, \gamma[x]] = 0$ to get (at $\langle \hat{x}, \gamma[\hat{x}] \rangle$)

$$0 = u_1 \cdot 1 + u_2 \cdot \gamma' = \langle u_1, u_2 \rangle \cdot \langle 1, \gamma' \rangle = u'[x, \gamma[x]] \cdot \langle 1, \gamma' \rangle$$

It follows that that the gradients f' and g' are orthogonal to $\langle 1, y'[\hat{x}] \rangle$ at the point $\langle \hat{x}, \hat{y} \rangle$. Therefore they are collinear. That is

$$f'[\hat{x}, \hat{y}] = \lambda\, g'[\hat{x}, \hat{y}] \tag{26.46}$$

for some number λ. These first-order necessary conditions provide two equations in three unknowns: x, y, and λ. A third equation is provided by the constraint: $g[x, y] = 0$. Together, these three equations provide the first-order necessary conditions for the constrained optimization.

Example 26.6 Let $f = \langle x, y\rangle \mapsto xy$ and let $g = \langle x, y\rangle \mapsto x + 2y - 8$, where the constraint is $g[x, y] == 0$. The $f' = \langle x, y\rangle \mapsto \langle y, x\rangle$ and $g' = \langle x, y\rangle \mapsto \langle 1, 2\rangle$. Imposing collinearity of the gradient and rewriting the constraint produces three equations in three unknowns:

$$y = \lambda \qquad x = 2\lambda \qquad x + 2y = 8$$

Solve these for the stationary point $\langle x, y, \lambda\rangle = \langle 4, 2, 2\rangle$. Note that $f'[4, 2] = \langle 2, 4\rangle$ which is indeed collinear with $g'[4, 2] = \langle 1, 2\rangle$.

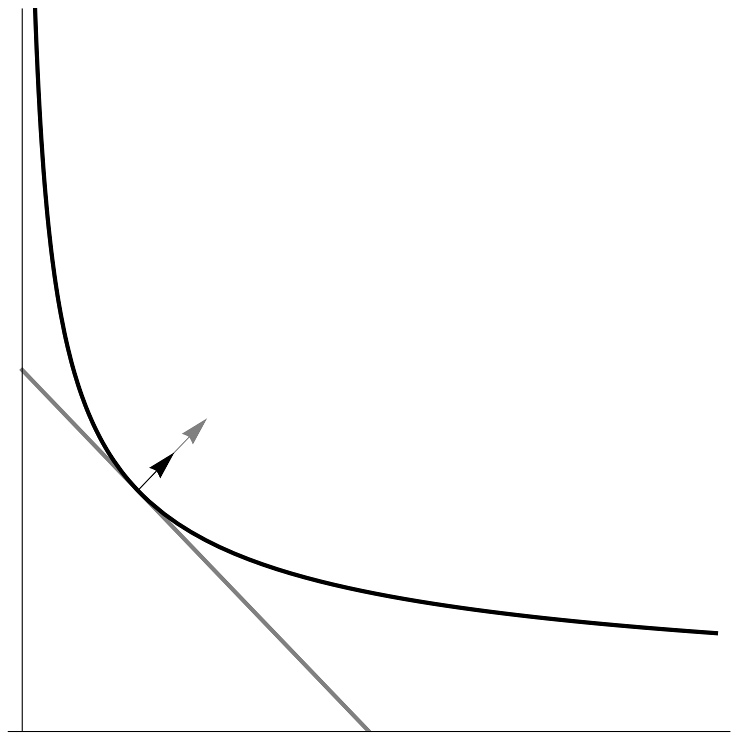

Figure 26.2: Colinearity of Gradients

26.3.2 Lagrangian Formulation

Lagrangian functions provide a convenient way to summarize the conditions for optimization discussed above. Let $\boldsymbol{x} = \langle x_1, \ldots, x_N\rangle$ and once again consider the problem of maximizing $f[\boldsymbol{x}]$ subject to $g[\boldsymbol{x}] = 0$, for real functions f and g define on $\mathbb{R}^N$. Write the Lagrangian function like this:

$$\mathcal{L} = \langle \boldsymbol{x}, \lambda\rangle \mapsto f[\boldsymbol{x}] - \lambda\, g[\boldsymbol{x}] \tag{26.47}$$

Note that this notation continues to conveniently represent all of the objective function arguments as $\boldsymbol{x}$.

To seek a stationary point with respect to all of the variables of $\mathcal{L}$, set its gradient to $\mathbf{0}$ and solve the resulting set of equations.

$$\begin{aligned} \nabla f[\boldsymbol{x}] - \lambda \nabla g[\boldsymbol{x}] &= \mathbf{0} \\ g[\boldsymbol{x}] &= 0 \end{aligned} \tag{26.48}$$

Example 26.7 Maximize xy subject to $x + y = 4$. As a convenience, we set up the Lagrangian:

$$\mathcal{L}[x, y, \lambda] = xy - \lambda(x + y - 4)$$

Take the first derivatives and set to zero:

$$\begin{aligned} \mathcal{L}_x = 0 &\implies y - \lambda = 0 \\ \mathcal{L}_y = 0 &\implies x - \lambda = 0 \\ \mathcal{L}_\lambda = 0 &\implies x + y - 4 = 0 \end{aligned}$$

We will refer to these as the first-order conditions of the problem. The first two imply $x = y$, we can plug this into the constraint to get $x = 2$, $y = 2$.

Exercise 26.1 Maximize $z = x^\alpha y^\alpha$ where $\alpha = .5$ and subject to $x + y = 4$. Use the method of substitution, the equal slope method and the technique of Lagrange multipliers. And show that all methods yield identical solutions.

Exercise 26.2 The principle of optimization plays a fundamental role in the microeconomic theory of consumer choice. The principle is precisely stated in a two-good case (x, y) as the maximization of a utility function $u[x, y]$ subject to a budget constraint $p_x x + p_y y = M$.

Suppose $u[x, y] = xy$, solve the consumer utility optimization problem:

$$\max u[x, y] = xy$$

subject to: $p_x x + p_y y = M$. Show that the optimal levels of x and y are given as $x = M/2p_x$ and $y = M/2p_y$. Hint: the solution to this problem is a straight forward application of the method of substitution or the method of Lagrange multipliers for constrained optimization problems.

Exercise 26.3 Often the structural equations we use in our comparative static experiments are derived from some underlying maximization behavior. For example, in the standard consumer maximization problem, the consumer maximizes $u[x, y]$ subject to $p_x x + p_y y \leq I$. Here x and y are two consumer goods, p_x and p_y are their prices, and I is the consumer's income. Letting λ be the Lagrange multiplier associated with the budget constraint and assuming an interior solution, the utility maximizing consumption bundle must satisfy the following three first order conditions:

$$U_x[x,y] - \lambda p_x = 0 \qquad U_y[x,y] - \lambda p_y = 0 \qquad p_x x + p_y y = I$$

Solve this three equation system for the effects of a change in consumer income on the consumption of x and y. Note: λ is also endogenous.

Example 26.8 Consider the function defined by the rule $f[x,y] = xy$. This clearly has no extreme points. Now minimize this function subject to the constraint that $x, y \geq 0$. The minimizer is clearly $\langle 0, 0 \rangle$. Note that the Hessian is indefinite at this point.

26.3.3 Second-Order Conditions

Naturally, the nature of a stationary point still reflects the curvature of the objective function. However, only the behavior on the constraint set is relevant. This means that we cannot simply examine the hessian of the objective function when thinking about the relevant curvature. For example, a constrained maximum need not be locally maximal, except on the constraint set.

One method to address this is by means of a ***bordered*** hessian. When there is a single equality constraint, this simply adds the gradient of the constraint as a border for the hessian of the Lagrangian.

$$\boldsymbol{H} = \begin{bmatrix} \mathbf{0} & g'[\boldsymbol{x}] \\ (g'[\boldsymbol{x}])^\top & D^2_{\boldsymbol{x}}\mathcal{L}[\boldsymbol{x}] \end{bmatrix} \tag{26.49}$$

When there are k equality constraints and $\boldsymbol{x}$ is an N-vector, we need to check the last $N-k$ leading principal minors. To ensure a maximum at $\hat{\boldsymbol{x}}$, we need $D^2_{\boldsymbol{x}}\mathcal{L}[\hat{\boldsymbol{x}}, \hat{\lambda}]$ to be negative definite on the constraint set. The determinant of the (order $(N+k)$) hessian matrix must be $(-1)^N$, and the signs of the last $N-k$ leading principal minors must alternate in sign. In the familiar case of 1 equality constraint and $\boldsymbol{x}$ is an 2-vector, we only need to check the last leading principal minor, and its sign must be positive.

Example 26.9 Return to the example $f = \langle x, y \rangle \mapsto xy$ and $g = \langle x, y \rangle \mapsto x + 2y - 8$, so that $\mathcal{L} = \langle x, y, \lambda \rangle \mapsto xy - \lambda(x + 2y - 8)$. Recall that we found a stationary point $\langle x, y, \lambda \rangle = \langle 4, 2, 2 \rangle$. Construct the bordered hessian

$$|\,\boldsymbol{H}| = \det \begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \end{bmatrix} = 4 \tag{26.50}$$

Example 26.10 Consider the consumer choice problem above. Recall that by substituting the constraint we produced the unary function

$$x \mapsto u[x, \gamma[x]]$$

with the first-order necessary condition at $\langle \hat{x}, \hat{x} \rangle$

$$u_x[x, \gamma[x]] + u_y[x, \gamma[x]]\gamma' = 0$$

This unary function is concave if

$$0 > u_{xx} + u_{x,y}\gamma' + u_{yx}\gamma' + u_{yy}(\gamma')^2$$

Next, consider the bordered hessian for the Lagrangian, $\mathcal{L} = u[x, y] - \lambda(p_x x + p_y y - w)$.

$$H = \begin{bmatrix} 0 & -p_x & -p_y \\ -p_x & u_{xx} & u_{xy} \\ -p_y & u_{yx} & u_{yy} \end{bmatrix}$$

Use the rule of Sarus to find the determinant:

$$|H| = p_x p_y(u_{x,y} + u_{yx}) - (p_y^2 u_{xx} + p_x^2 u_{yy})$$

Factoring out $-p_y^2$, get

$$|H| = -p_y^2 \left(u_{xx} + (p_x/p_y)^2 u_{yy} - (p_x/p_y)(u_{x,y} + u_{yx})\right)$$

Recalling $\gamma' = -p_x/p_y$, the hessian determinant is positive iff and only iff the unary condition for concavity is satisfied.

26.4 Generalizing: Convex Optimization with Binding Constraints

This section focuses on the maximization problem. (Minimization of f is maximization of $-f$.) The basic problem is to maximize the continuously differentiable concave function $f : \mathbb{R}^N \to \mathbb{R}$ subject to being in a convex constraint set.

$$\max_{x \in C} f[x] \tag{26.51}$$

Here C is is a non-empty convex set determined by the convex and continuously differentiable constraint functions g_i as follows:

$$g_i[x] \leq 0 \qquad i \in [1 .. I] \tag{26.52}$$

The feasible set comprises the points that satisfy all the constraints. A problem may have no feasible points; the constraints may be inconsistent. If the feasible set is nonempty, then it is convex. (It is the intersection of the subgraphs of the g_i, and the subgraphs of a convex function are convex.) A solution is a maximizer in the set of feasible points. (If the feasible set is unbounded, there may be no maximizer.)

Note that a local optimum for this problem is a global optimum. To see this, assume that x_ℓ is a local optimizer in the neighborhood $\mathrm{H}(x_\ell$. For any point $x \notin \mathrm{H}[x_\ell]$, we can form a strict convex combination of x_ℓ and x with enough weight λ on x_ℓ to ensure we are inside $\mathrm{H}[x_\ell]$. If $f[x] > f[x_\ell]$, then concavity of f would then imply that $f\left(\lambda x_\ell + (1-\lambda)x\right) > f[x_\ell]$, violating the assumption that x_ℓ was a local optimum. So we must have $f[x] \leq f[x_\ell]$, implying x_ℓ is a global maximizer.

Looking forward, after additional investigation a solution can be characterized by the following. At the solution value, the gradient of the objective function is a linear combination of the gradients of the binding constraints. The weights of this linear combination are called ***Lagrange multipliers***, and the multipliers for the nonlinear constraints are nonnegative.

Recall that the lower contour sets of convex functions are convex, while the upper contour sets of concave functions are convex.

Describe a level set by $k = f[\boldsymbol{x}]$. Assume f is smooth. Let $\boldsymbol{x}[t]$ be a curve on this level set, so $\nabla f \cdot \mathrm{d}\boldsymbol{x}/\mathrm{d}t = 0$. That is, the gradient is perpendicular to the tangent to any curve on the surface.

The key intuition of an optimum is that any movement you can make away from the optimum is to your detriment. If the optimization problem is subject to constraints, the movements you can make may of course be limited by these constraints. Our characterization of the optimum turns on this idea. A constraint is ***binding*** if it stops you from moving in a direction that would be beneficial. E.g., a budget constraint is biding if it prevents you from buying more utility-improving goods and services. This means that the direction of increase of the objective function is in a direction that is infeasible due to the constraint. This means that any movement in that direction would cause a constraint violation. That is, any movement in a direction of increase for the objective function would cause a disallowed increase in the value of the constraint function. So any of increase of the objective function must be a direction of increase for the constraint function. This means the gradients point in the same direction.

Recall that a vector $\boldsymbol{h}$ is a direction of increase of f at $\boldsymbol{x}$ if it is in the positive half-space of ∇f. That is, $\boldsymbol{h} \cdot \nabla f[\boldsymbol{x}] > 0$. So if we have an optimum where the gradient is nonzero, there must not be an feasible points in the positive half-plane generated by the gradient of the objective function. A traditional formulation of the two-good consumption problem illustrates this, as in Figure 26.3.

If we add a second constraint, we cannot ask that the gradient of the two constraints and the gradient of the objective function all point in the same direction. Instead we require that the gradient of the objective function be a conical combination of the constraint function gradients. This is illustrated by Figure 26.4. Here we require not only that the budget constraint be satisfied, but that the consumer receive a minimum amount of the first good.