Testing the protein-leverage hypothesis using population surveillance data

Alistair M. Senior1,2,3, David Raubenheimer1,2 and Stephen J. Simpson1,2

1Charles Perkins Centre, 2School of Life and Environmental Sciences, and 3School of Mathematics and Statistics, The University of Sydney, Camperdown, NSW 2006, Australia

AMS, 0000-0001-9805-7280; DR, 0000-0001-9050-1447

It is hypothesized that humans exhibit ‘protein leverage’ (PL), whereby regulation of absolute protein intake results in the over-consumption of non-protein food on low percentage protein diets. Testing for PL using dietary surveillance data involves seeking evidence for a negative association between total energy intake and percentage energy from protein. However, it is unclear whether such an association might emerge without PL due to the structure of intake data (protein and non-protein intakes have different means and variances and covary). We derive a set of models that describe the association between the expected estimate of PL and the distributions of protein and non-protein intake. Models were validated via simulation. Patterns consistent with PL will not emerge simply because protein intake has a lower mean and/or variance than non-protein. Rather, evidence of PL is observed where protein has a lower index of dispersion (variance/mean) than non-protein intake. Reciprocally, the stronger PL is the lower the index of dispersion for protein intake becomes. Disentangling causality is ultimately beyond the power of observational data alone. However, we show that one can correct for confounders (e.g. age) in generating signals of PL, and describe independent measures that can anchor inferences around the role of PL.

1. Introduction

The way that an organism regulates its intake of nutrients is a core question in metabolic, ecological and evolutionary science. Through the twentieth century, evolutionary ecology was heavily focused on the assumption that the organism is adapted to maximize a single currency, usually energy intake, per unit of time [1]. This focus on energy overlooks the fact that different nutrients have different physiological roles. Working in parallel over this period, medical nutrition science looked beyond
energy, to consider how deficits of specific nutrients such as protein and other micronutrients might be affecting health [2]. Medical science, however, paid less attention to the adaptive basis for appetites, and how diseases could be a by-product of the interaction between appetites and the nutritional environment [3].

More recently, the geometric framework for nutrition (GFN) has emerged from evolutionary ecology and is being now used in human nutrition [4]. The GFN combines and builds on the preceding perspectives to understand how appetites have evolved to regulate the intake of multiple nutritional dimensions simultaneously. The GFN adopts a state-space approach, where the intake of \( n \) nutrients is considered in an \( n \)-dimensional space (figure 1a). A core concept is the intake target. The intake target is a coordinate or region within the space representing the intake of the \( n \)-nutrients that an organism seeks to achieve by eating the foods available (figure 1a). Foods are represented by vectors or ‘nutritional rails’ that project through the space from the origin with slope determined by the ratio of nutrients therein (figure 1b). Integrated over a longer timeframe a rail could represent the ratio of nutrients in an organism’s diet rather than a single food.

Given ad libitum access to a food/diet with a nutritional ratio that does not match that of the intake target, an organism must compromise between over- and under-consuming different nutrients. One strategy is to strictly regulate the intake of one nutrient at the expense of under/over-consuming others (figure 1c). Such a strategy may have important consequences for total energy intake.
For example, consider the case where an organism strictly defends its intake of protein energy at the expense of non-protein energy (figure 1c). Here, total energy, which is the sum of protein and non-protein intake achieved on a food rail, will fall as the proportion of protein in the rail increases (figure 1d); this is the concept of ‘protein leverage’ (PL; [5,6]). The relationship between energy intake and protein content in a food/diet under protein leverage can be described as

\[ E = Pp^L, \]  

where \( E \) is total energy intake, \( p \) is the proportion of energy contributed by protein in the diet, \( P \) is a constant and \( L \) is the strength of leverage [7,8]. Where the protein target is defended strictly (e.g. figure 1c), \( L = -1 \) and \( P \) will be equal to the organism’s target intake of protein as well as the absolute amount of protein eaten (figure 1d). As the protein target becomes less strictly defended, \( L \) will increase and \( P \) will deviate from the protein target; \( L = 0 \) suggests that the protein and non-protein energy are regulated equally strongly and there will be no association between \( p \) and \( E \) (i.e. no protein leverage).

Evidence for PL has been observed across numerous animal species in controlled laboratory experiments (see [4,8]). Detailed field studies on non-human primates in which the feeding behaviour of individuals was followed intensively over extended periods have also documented patterns of macronutrient intake consistent with PL (e.g. [9,10]). In addition to such behavioural evidence for PL, laboratory studies have made accelerating progress toward understanding the mechanisms that control protein appetite, both in mammals and invertebrate models (e.g. [11–13]).

The degree to which humans exhibit PL is a critical aspect of our biology with consequences for non-communicable disease. The protein leverage hypothesis (PLH) posits that PL in humans coupled with the dilution of dietary protein in the food supply has contributed to the global obesity epidemic [6,8]. There are challenges in translating inevitably simplified experimental designs conducted under somewhat artificial conditions to the behaviour of free-living humans [14,15]. Nonetheless, randomized controlled trials have demonstrated a causal effect of protein content on food intake in a manner entirely consistent with PL [5,16,17]. While two trials did not find elevated energy intake on a 5% protein diet [18,19], these studies must be interpreted with caution as the levels of protein tested are physiologically unsustainable (see [8]).

Detecting PL in a laboratory setting does not, however, prove that it plays a role in driving energy hyperphagia and obesity in realistic settings. It is therefore important to also test for PL in population data, for example from diet surveys and cohort studies (e.g. [20–22]). These data are typically gathered from surveys, questionnaires, food diaries and related dietary assessment tools. From the pre-existing among-individual variation in the proportion of protein in the diet and total energy intake in these data, one may then make an estimate of the strength of PL. This can be done by directly fitting equation (1.1) and taking the \( L \) coefficient. Alternatively, in his derivation Hall [7] shows that equation (1.1) can be expressed as

\[ \log(E) = \log(P) + L \log(p), \]  

which is useful in an applied context as it allows one to test for PL using a standard parametric linear regression of the form

\[ y_i = \alpha + \beta x_i + e_i, \]  

where \( y_i \) is the log total intake of the \( i \)th observation in a dataset, \( \alpha = \log(P) \), \( \beta = L \) (\( \alpha \) and \( \beta \) are coefficients estimated from the data), \( x_i = \log(p) \) for the \( i \)th observation (i.e. the predictor data), and \( e_i \) is the residual for the \( i \)th observation. This approach is helpful because linear regression is easily implemented, widely understood and provides a single measure of the strength of leverage (\( \beta = L \)) that can be compared across studies and with theoretical expectations. Equation (1.3) is particularly useful in an epidemiological context as one can statistically correct for potential confounders (e.g. age or sex) by adding covariates in the form of a multiple regression. Using population data and the approaches suggested above, for example, Saner et al. [20] estimate the strength of PL in a cohort of younger people with obesity as \( L = -0.48 \). While not complete (i.e. \( L > -1 \)) this degree of leverage is strong enough to have played a significant role in the obesity epidemic [7]. We note here that a non-parametric regression could also be implemented to assess the degree to which \( x_i \) and \( y_i \) negatively covary (i.e. a signal of PL). However, unlike parametric regression, such models do not provide a succinct and widely comparable single coefficient of the strength of PL, and are not as widely used by epidemiologists.

A question that arises from the estimation of PL using population data is: will a negative association between energy intake and percentage energy from protein arise spuriously from the relative
distributions of protein and non-protein energy intake? The answer is not necessarily obvious because: (i) at the population level, protein and non-protein energy tend to be correlated owing to among-individual variation in factors affecting total food intake, (ii) protein and non-protein energy may have different variances and (iii) protein naturally makes up a relatively small proportion of the diet. Here we present a series of theoretical models that examine the expected associations between the strength of PL and population-level statistics on nutrient intake in the absence and presence of true PL. These models are validated using statistical simulation, where we also challenge underlying assumptions. Finally, we discuss how one might holistically assess evidence for/against PL in a population context.

1.1. Model 1

All calculations and simulations were performed using R v. 4.1.0 [23], and code is publicly available. Code is maintained at https://github.com/AlistairMcNairSenior/PL_Theory, and an archived version has been released [24].

We begin by deriving an expression for the expected value of $L$ as estimated by equation (1.3) as a function of the means and variances of protein and non-protein intake and their covariance. It is critical to note that in this model the explicit assumption that PL is acting is not the starting point. Rather our aim is to see how the value of $L$ might be affected by the relative distributions of protein and non-protein intake.

A model of intake that does not begin with the assumption of PL would be one in which protein intake is a random variable $U$, and non-protein intake is a random variable $V$ and the two have no explicit causal effect on one another. We can model these intakes as a bivariate normal distribution with means $\mu_U$ and $\mu_V$ standard deviations $\sigma_U$ and $\sigma_V$ and covariance $\sigma_UV$ (the correlation between $U$ and $V$ will be $\rho_UV = \sigma_UV/(\sigma_U\sigma_V)$). Positive covariances are expected where there are factors within the population (beyond PL) that generate variance in total food intake (e.g. activity level). We revisit the assumption of normality below.

In this case total energy intake would be the random variable $Z = U + V$. Because the sum of normal distributions is itself normal, $Z$ will be normally distributed with mean and standard deviation

$$
\mu_Z = \mu_U + \mu_V, \\
\sigma_Z = \sqrt{\sigma_U^2 + \sigma_V^2 + 2\rho_UV\sigma_U\sigma_V},
$$

and $Z$ (total energy intake) and $U$ (intake of protein energy) will have covariance, $\sigma_UZ$, and correlation, $\rho_UZ$

$$
\sigma_UZ = \sigma_{U(U+V)} = \sigma_UU + \sigma_UV = \sigma_U^2 + \sigma_UV, \\
\rho_UZ = \frac{\sigma_UZ}{\sigma_U\sigma_Z}.
$$

Let the proportion of total energy coming from protein be random variable $W = U/Z$. $W$ is the ratio of two non-central correlated random variables and will therefore, following equation (1.13) in Pham-Gia et al. [25], have the probability density function

$$
f_W(w) = K \left( 2 \left( 1 - \rho_U^2 \right) \sigma_U^2 \sigma_Z^2 \right)^{-1/2} \exp \left[ -\frac{\rho_U^2 \mu_U^2 - \sigma_UZ \mu_U \mu_Z + \rho_U \sigma_U^2 \sigma_Z^2}{2(1 - \rho_U^2)\sigma_U \sigma_Z} \right],
$$

where

$$
K = \frac{1}{2\pi \sigma_U \sigma_Z \sqrt{1 - \rho_U^2}} \exp \left[ \frac{-\rho_U^2 \mu_U^2}{2(1 - \rho_U^2)\sigma_U^2} \right],
$$

$$
f_r(w) = \frac{\rho_U^2 \sigma_U^2 w + 2\sigma_UZ(\mu_U w + \mu_Z) - \rho_U \sigma_U^2 \sigma_Z^2}{2\sigma_U^2 \sigma_Z^2 (1 - \rho_U^2)\sigma_U^2 \sigma_Z^2},
$$

$$
f_r(n, m) = \frac{\Theta(n, m)}{\Gamma(n)\Gamma(m)} x^n y^m,
$$

where equation (1.11) is Kummer’s classical confluent hypergeometric function of first kind, which we have estimated computationally using the ‘kummerM’ function in the fAsianOptions package in R [26].
The mean, or expected value, of $W$ (proportion of energy coming from protein), and its variance will be

$$\mu_W = \mathbb{E}(W) = \int w f_W(w) \, dw,$$

$$\sigma_W^2 = \int (w - \mu_W)^2 f_W(w) \, dw,$$

(1.12)

(1.13)

and the covariance, and correlation, between $Z$ (total energy) and $W$, will be

$$\sigma_{WZ} = \mathbb{E}(WZ) - \mathbb{E}(W)\mathbb{E}(Z) = \mathbb{E}(U) - \mathbb{E}(W)\mathbb{E}(Z) = \mu_U - \mu_W\mu_Z,$$

$$\rho_{WZ} = \frac{\sigma_{WZ}}{\sigma_W\sigma_Z}.$$

(1.14)

(1.15)

Here we see that the proportion of energy coming from protein ($W$) and total energy intake ($Z$) will negatively covary (a sign of PL) when the mean of protein intake ($U$) is less than the mean of proportion of energy coming from protein multiplied by the mean of total energy intake.

Estimating the strength of leverage via equation (1.3) requires fitting $W$ and $Z$ on the log scale. Thus, to get the expected value of $L$, which is the slope from equation (1.3), requires us to find the probability distribution function for the log of $W$. Let the log of $W$ be random variable $X$, then the distribution function for $X$, $f_X(x)$, using the method of transformation is

$$h(w) = \log (w) = x,$$

$$h^{-1}(x) = e^x = w,$$

$$\frac{dh^{-1}}{dx} = e^x,$$

$$f_X(x) = f_W(h^{-1}(x)) \frac{dh^{-1}}{dx} = f_W(e^x) e^x.$$

(1.16)

(1.17)

(1.18)

(1.19)

The mean and variance of $X$ (log proportion energy from protein) can be calculated as

$$\mu_X = \mathbb{E}(X) = \int x f_X(x) \, dx,$$

$$\sigma_X^2 = \int (x - \mu_X)^2 f_X(x) \, dx.$$

(1.20)

(1.21)

Now treating the log $Z$ (total energy intake) as random variable $Y$, a commonly used approximation for the covariance between $X$ and $Y$ based on the Taylor series expansion is

$$\sigma_{XY} = \log \left(1 + \frac{\sigma_{WZ}}{\mu_W\mu_Z}\right) = \log \left(1 + \frac{\mu_U - \mu_W\mu_Z}{\mu_W\mu_Z}\right).$$

(1.22)

Finally, based on this covariance the expected slope of equation (1.3), $\beta$, which is equivalent to $L$ in equations (1.1) and (1.2), can be approximated from population means and variances as

$$E(\beta) = \frac{\sigma_{XY}}{\sigma_X^2} = \frac{\log (1 + ((\mu_U - \mu_W\mu_Z)/\mu_W\mu_Z))}{\sigma_X^2} = E(L).$$

(1.23)

1.2. Model 1: numerical results and simulation

Figure 2a shows the expected strength of leverage, $L$, from equation (1.23) as a function of the relative contribution of protein ($\mu_U$) to total energy intake ($\mu_U + \mu_V$) assuming $\mu_U + \mu_V = 8700$, which is the recommended average adult daily intake in kJ used by Food Standards Australia & New Zealand, and the correlation between protein and non-protein intake ($\rho_{UV}$). Here we see that as long as protein contributes less than 50% of total energy on average, we expect to see positive values of $L$ in the absence of protein leverage (PL). This analysis indicates that the fact that protein makes up the smaller fraction of total energy, in and of itself, is not enough to drive a spurious correlation that is consistent with the presence of PL (i.e. negative values of $L$).

The analysis in figure 2a holds the variance in protein ($U$) and non-protein ($V$) intake constant as the means of each change. However, for many types of data (e.g. lognormally distributed data) there tends to be a positive association between the mean and the variance. Thus, we might expect protein to have lower variance in intake than non-protein if it has a lower mean. We therefore repeated these analyses
holding the index of dispersion (ID: variance/mean) for the two nutrients constant. Where there was no correlation between $U$ and $V$, $L$ was constant and zero regardless of the contribution of protein to total intake (figure 2b). Where there is a stronger correlation (and the ID is constant) there are positive values of $L$ when $U$ tends to make the smaller contribution to energy intake (figure 2b). These results suggest that evidence for PL (i.e. $L < 0$) will not arise simply because protein makes up the lower percentage of total energy and has a commensurately lower variance.

Finally, one might speculate on the value of $L$ when protein has a lower mean intake and a variance even lower than that expected given the mean (i.e. ID$_U$ < ID$_V$). Figure 3 shows $L$ as a function of the relative contribution of $U$ to energy, the relative IDs of $U$ and $V$, and the correlation there-between ($\rho_{UV}$). Where $\rho_{UV} = 0$, apparent indication for PL ($L < 0$) can be expected when the ID for $U$ is less than that for $V$ (figure 3); put another way, when protein intake has lower than expected variance given its mean, evidence suggesting PL can be observed. As the correlation between protein and non-protein intake increases, negative values of $L$ are only estimated when ID$_U$ is substantially smaller than ID$_V$ and/or protein approaches 50% of total energy.

To test the performance of equation (1.23), we ran a series of statistical simulations. These simulations allow us to evaluate the accuracy of approximations and challenge assumptions, made in the derivation of equation (1.23). For each of the conditions shown in figure 3, we simulated 20,000 protein ($U$) and non-protein ($V$) intakes from a bivariate random-normal distribution using the mvrnorm function in the package MASS [27]. From these simulated data we calculated (natural) log total intake ($Y$) and (natural) log proportion of energy from protein ($X$), fitted the model shown in equation (1.3) using the ‘lm’ function in base R and extracted the relevant coefficient. The top row of figure 4 shows the value of $L$ under each condition as estimated using equation (1.23), and as estimated from simulated data; there is a very strong positive correlation indicating that equation (1.23) performs well. Our approximation and the simulated data match near-perfectly for more negative values of $L$. For more positive values of $L$, equation (1.23) slightly underestimates $L$ relative to simulated values (figure 4, top row).

We simulated our intake data from a normal distribution. Because normal distributions are not bound at zero (i.e. can span zero) simulations from them can generate negative values, which become more abundant as the variance increases. This is problematic in the current model because negative values fail in equation (1.3) as log($\leq$0) is undefined, and consequently any negative simulated values were excluded from our analyses. Furthermore, true intakes of protein and non-protein are ratio scale and cannot be negative (i.e. one cannot eat less than 0 kJ). To test the sensitivity of our results to assumptions of normality we repeated our simulation but using a lognormal distribution (note a lognormal distribution is not the same as the log of a normal distribution; see [28]). Lognormal distributions are useful in the current context because much like real intake data they are bounded at

![Figure 2](image-source)
zero (i.e. cannot be negative). Lognormal data were simulated by taking the exponential of data from 'mvrnorm' where \( \rho_{UV} = 0 \) and using 'SimCorRVs' in the anySim package \[29\] where \( \rho_{UV} \neq 0 \). There was a tight association between simulated values of \( L \) and those from equation (1.23). However, where \( \rho_{UV} \neq 0 \) simulated values were greater than those from equation (1.23) where \( L \) was positive (figure 4, bottom row). It is important to note that, while equation (1.23) does underestimate \( L > 0 \), equation (1.23) is near-perfect when \( L \) is negative, which is the primary area of concern when inferring the presence of PL.

In summary, model 1 indicates that when protein makes up the smaller component of the diet one may detect a pattern consistent with PL only when protein intake has lower-than-expected variance given the mean (figure 3). However, we now demonstrate that such an association need not simply be spurious. Rather, PL is a mechanism by which exactly these distributions are generated.

1.3. Model 2

That PL will generate lower variance in protein intake than in non-protein intake can be understood intuitively; PL occurs because protein intake is more strongly regulated than non-protein intake. However, we now present a model that explicitly assumes PL and work up from there to assess how the strength of leverage will affect the relative distributions of protein and non-protein intake.

Taking equation (1.3) as a model of PL, we can begin by assuming that the proportion of energy from protein is a random variable, \( W \), that follows a beta distribution, and will then have probability density function
density function

\[ f_W(w) = \frac{\Gamma(\kappa + \tau)}{\Gamma(\kappa)\Gamma(\tau)} w^{\kappa-1}(1-w)^{\tau-1}, \quad (1.24) \]

where \( \kappa \) and \( \tau \) are the shape parameters. We note that the mean and variance for \( W \) can be calculated directly from \( \kappa \) and \( \tau \) as \( \text{E}(W) = \kappa/(\kappa + \tau) \) and \( \text{Var}(W) = (\kappa \cdot \tau)/((\kappa + \tau)^2+(\kappa + \tau + 1)) \). Also, we note that the probability density function for the proportion of energy coming from non-protein, \( Q = 1 - W \), will be simply

\[ f_Q(q) = f_W(1 - q). \quad (1.25) \]

The log proportion energy from protein, \( X = \log(W) \) as is used in equation (1.3) will, following the method of transformation (as in equations (1.16)–(1.19)), have the probability density function \( f_X(x) \), mean \( (\mu_X) \) and variance \( (\sigma_X^2) \)

\[ f_X(x) = f_W(e^x)e^x, \quad (1.26) \]
\[ \mu_X = E(X) = \int xf_X(x) \, dx, \quad (1.27) \]
\[ \sigma_X^2 = \int (x - \mu_X)^2f_X(x) \, dx. \quad (1.28) \]

Based on equation (1.3) the probability density function for the log of total energy intake, \( Y \), will then be the convolution of \( f_X \) adjusted for \( \beta \) (which is the strength of leverage, \( L \)) and the probability density function for a random normal distribution \( R \) with mean \( = \alpha \) and with (residual) variance \( \sigma_e^2 \)

\[ f_R(r) = \frac{1}{\sigma_e \sqrt{2\pi}} \exp \left[ - \left( \frac{1}{2\sigma_e^2} \right) (r - \alpha)^2 \right], \quad (1.29) \]
\[ f_Y(y) = \int f_X(x) \left| \frac{1}{\beta} \right| f_R(y - x) \, dx. \quad (1.30) \]
The mean and variance of $Y$ will be
\[
\begin{align*}
\mu_Y &= E(Y) = \int y f_Y(y) \, dy, \\
\sigma_Y^2 &= \int (y - \mu_Y)^2 f_Y(y) \, dy.
\end{align*}
\tag{1.31}
\]

We can then apply the method of transformation to $Y$ to get the distribution, mean and variance in energy intake on the natural scale, $Z$ as
\[
\begin{align*}
 f_Z(z) &= f_Y(\log z) \frac{d(\log z)}{dz} = f_Y(\log z) \frac{1}{z}, \\
\mu_Z &= \int z f_Z(z), \\
\sigma_Z^2 &= \int (z - \mu_Z)^2 f_Z(z).
\end{align*}
\tag{1.33}
\tag{1.34}
\tag{1.35}
\]

We can approximate the mean and variance in log absolute protein intake ($\log(U)$) as:
\[
\begin{align*}
\mu_{\log U} &= \mu_X + \mu_Y, \\
\sigma_{\log U}^2 &= \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY}, \\
\sigma_{XY} &= \beta \sigma^2_X.
\end{align*}
\tag{1.36}
\tag{1.37}
\tag{1.38}
\]

which back-transforms from the log scale to give expected value (mean) and variance for absolute protein intake, $U$, as
\[
\begin{align*}
\mu_U &= \exp(\mu_{\log U} + \frac{1}{2} \sigma_{\log U}^2), \\
\sigma_U^2 &= (e^{\sigma_{\log U}^2} - 1)\mu_U.
\end{align*}
\tag{1.39}
\tag{1.40}
\]

Finally, we can find the mean and variance in non-protein energy intake, $V$. The mean is straightforwardly
\[
\mu_V = \mu_Z - \mu_U. 
\tag{1.41}
\]

The variance of $V$ can be approximated on the log scale (and back-transformed via equation (1.40)) as
\[
\sigma_{\log V}^2 = \sigma_Y^2 + \sigma_V^2 + 2\sigma_{SY}, 
\tag{1.42}
\]
where, $\sigma_{SY}$ is the covariance between log total energy intake ($\log(Z) = Y$) and the log proportion of the energy from non-protein ($S = \log(1 - W)$). The covariance $\sigma_{SY}$ can be approximated as
\[
\begin{align*}
\sigma_{(1-W)Z} &= -\sigma_{WZ} = -(E(ZW) - E(Z)E(W)) = -(E(U) - E(Z)E(W)) = -(\mu_U - \mu_Z\mu_W), \\
\sigma_{SY} &= \sigma_{\log(1-W)\log Z} = \log\left(1 + \frac{\sigma_{(1-W)Z}}{E(1 - W)\mu_Z}\right) = \log\left(1 + \frac{(\mu_U - \mu_Z\mu_W)}{(1 - \mu_W)\mu_Z}\right),
\end{align*}
\tag{1.43}
\tag{1.44}
\]

The variance in $S$ (log proportion of energy from non-protein) can be found as
\[
\begin{align*}
f_S(s) &= f_Z(ws) \exp(s), \\
\mu_S &= E(S) = \int s f_S(s) \, ds, \\
\sigma_S^2 &= \int (s - \mu_S)^2 f_S(s) \, ds.
\end{align*}
\tag{1.45}
\tag{1.46}
\tag{1.47}
\]

1.4. Model 2: numerical results and simulation

We next calculated the index of dispersion ratio (IDR = $ID_U/ID_V$) for protein ($U$) and non-protein ($V$) intake as a function of the strength of leverage ($L$), using model 2 to find the relevant means and variances. Figure 5 shows IDR as a function of $L$ and the levels of variance in the proportion of energy from protein ($\sigma_{W}$), and assuming protein on average makes up 15%, 20% and 30% of energy. As $L$ increases toward 0 (i.e. PL gets weaker) the IDR trends toward 1 (i.e. equal IDs for protein and non-protein). This effect was most marked where we assumed there was higher among-individual variance in proportion of energy from protein ($\sigma_{W} = 0.07$) and protein made up a smaller proportion
of total energy ($\mu_W = 0.15$). These results show that stronger leverage will reduce the (mean-corrected) variance in protein intake relative to that for non-protein intake.

We performed a series of simulations to validate model 2. For each set of parameters shown in figure 5 we simulated 100k values of $W$ (proportion of energy from protein) from a beta distribution (‘rbeta’ function in base R) and 100k residual values from a normal distribution (‘rnorm’ function in base R) with mean of 0 and a variance of $\sigma^2_W$. We then calculated log total energy intake ($Y$) assuming PL via equation (1.3), and absolute total energy intake as $Z = e^Y$. Finally, values of absolute intake of protein and non-protein were calculated as $U = ZW$ and $V = Z(1-W)$, respectively. The means and variances of $U$ and $V$ were then used to calculate the simulated IDR. The approximated values of IDR from model 2 and those simulated were strongly correlated (figure 6). However, those from model 2 tended to be slightly higher than the simulated values where $\mu_W$ was low and the IDR was also closer to 1 (figure 6). Given this potential discrepancy, we have remade figure 5 using the simulated values of IDR. The trend regarding the effect of $L$ on the IDR was the same when calculated via model 2 and via simulation (figure 5 versus figure 7), hence both the analytical model and the simulation yield the same conclusions about the effects of protein leverage on the relative distributions of protein and non-protein intake.

In summary, model 2 shows that PL will lead to a lower index of dispersion (i.e. variance/mean) for protein than non-protein intake.
Figure 6. The ratio of the indices of dispersion for protein ($U$) and non-protein ($V$) intake ($\text{IDR} = \frac{\sigma_U^2/\mu_U}{\sigma_V^2/\mu_V}$) as approximated by model 2, and by simulation, for all parameter values shown in figure 4.

Figure 7. The ratio of the indices of dispersion for protein ($U$) and non-protein ($V$) intake ($\text{IDR} = \frac{\sigma_U^2/\mu_U}{\sigma_V^2/\mu_V}$) as a function of the strength of leverage ($L$) and the standard deviation in proportion of energy from protein ($\sigma_W$). Different values for the mean proportion of energy from protein ($\mu_W$) are assumed across different panels. Simulation was used to estimate the values of $\sigma_U^2$, $\mu_U$, $\sigma_V^2$, and $\mu_V$. The $\alpha$ value for equation (1.3) was fixed at $8700/\mu_W$ in all cases meaning that the modal total intake will be 8700, and residual variance, $\sigma_e^2$, was fixed at log(8700) $\times$ 0.02.
2. Discussion

A negative association between the proportion of dietary energy from protein and total energy intake is expected based on protein leverage (PL). This relationship has been reported in controlled experimental settings, where such an association cannot be explained as a numerical artefact [16]. Our modelling has shown that in analyses of population data such an association will not manifest simply because protein makes up the minor part of total energy intake, nor because protein has a lower variance in intake than non-protein energy (NP). Rather, to produce patterns consistent with PL requires that protein has a variance lower than NP after accounting for the difference in the mean. Human populations typically demonstrate these attributes, with protein making up around 15% of total energy intake, and being less variable around the mean than fat and carbohydrate intake (e.g. [30–32]). As we have shown, PL offers a mechanistic explanation for such patterns; however, the question of causality remains unanswered; do we see what appears to be PL because the variance in protein is lower than that for NP, or vice versa?

It is well established that causality cannot be attributed using population intake data alone (e.g. [33]); hence, the possibility will always remain that some factor other than PL disproportionately reduces the observed variance in protein relative to NP for these types of studies. We can, however, consider what such factors might be.

One possibility is that dietary assessment tools (DAT, e.g. dietary recall, food frequency questionnaires and related online tools) measure protein with higher precision (as distinct from accuracy) than NP, thereby reducing error and variance in protein without the involvement of PL. If this were the case, we would expect the variance in protein to be relatively similar whether measured by objective biomarkers (e.g. urinary nitrogen) or DAT. By contrast, the variance in total energy intake should be reduced when estimated using objective biomarkers such as the doubly labelled water method versus DAT. We tested these predictions using data presented in seven studies which compared biomarker- and DAT-derived values [34–40]. From these studies we have compared the sensitivity of variation in intakes to the method, using the coefficient of variation ratio (CVR: \( \frac{CV_{\text{Biomarkers}}}{CV_{\text{DAT}}} \)). The mean CVRs for total energy and protein intake are identical (mean CVR ± s.d., total energy = 0.96 ± 0.09, protein = 0.96 ± 0.13). This, albeit non-exhaustive, survey of the literature suggests that protein intake is measured with a similar degree of precision to total energy intake.

Another possibility is that the difference in variance between NP and protein emerges from heterogeneity in energy requirements among subjects, rather than PL. It is well established that variation in levels of physical activity or changes in basal metabolic rate with age generate variation in energy requirements [41]. Any cross-sectional analysis testing for PL relies on the observed variation in percentage energy from protein representing a ‘natural experiment’; it is assumed that any non-dietary drivers of variance in reported energy intake (e.g. sex, age or socio-economic status or misreporting) are randomly distributed or statistically controlled across the range of percentage protein values within the data. The risk of this kind of confounding is not confined to PL, of course, and epidemiology as a field is well versed in testing for these issues in observational data. In particular, many methods are available for identifying potential under/over-reporting of intakes [42]. The beauty of equation (1.2) as proposed by Hall [7] is that it is readily implemented as a linear regression. This means potential confounders can be corrected using the standard multiple linear regression approaches that are common to all epidemiologists (e.g. adding age or sex, or a misreporting index as a covariate in equation (1.3)), or through stratification and re-analysis (e.g. after removing implausible reporting [42]).

Having corrected for confounders, one approach to give confidence in a causal role for PL is to demonstrate variance in the food environments experienced by individuals within populations. This could be done, for example, by demonstrating variation in the proportions of different categories of foodstuffs eaten. Martínez Steele et al. [43] parsed NP and protein from the US National Health and Nutrition Examination Survey (NHANES) data into quintiles for ultra-processed food and beverage consumption (UPF; NOVA Category 4). The NOVA classification system is based on the extent of industrial processing [44] rather than the nutrient compositions of foods and beverages per se [45]. This approach is helpful as it provides a classifier one-step removed from the measurement of protein and NP themselves. As predicted by PL, Martínez Steele et al. [43] found that protein intakes were consistent across quintiles of UPF intakes, whereas total energy intakes varied greatly. These results in turn accords with those from a randomized clinical trial [46].

Support for a causal role of PL in affecting total energy intake (as opposed to PL arising spuriously from measurement errors) can also be established if other independently derived objective outcomes...
known to be affected by total energy intake conform to a priori expectations. For example, if lower percentage protein diets truly result in elevated energy intake, higher BMI would be expected on these diets. Other outcomes that might anchor the inference of causality of PL are biomarkers of diet itself, waist–hip ratio, percentage fat mass from DEXA and cardiometabolic markers. Observing that percentage protein shares a negative association with both adiposity and total energy would provide support for the PLH [6,8]. Of course, when tested with population data these more objective outcomes may have their own confounders that must be considered. Three particular issues of note are: (i) the positive association between BMI and under-reporting of non-protein energy intake, which would dampen signal of the PLH [47–49]; (ii) allometries and other complexities of growth and development in younger cohorts [20]; and (iii) if testing for an association with relatively static phenotypes (e.g. BMI or waist-hip ratio) ensuring that the observed nutrient intakes are representative of habitual diet (as opposed to a single measurement period).

Finally, it is worth acknowledging that our models and suggested analyses have been framed around the effects of protein in the diet on energy intake. This is because our work, and that by others, suggests that a two-dimensional model segregating protein and non-protein energy explains a significant amount of variation in food intake and captures features of energy intake overlooked by simpler models [32,50]. However, our insights are readily applicable to any nutritional component that is hypothesized to negatively feedback onto (i.e. leverages) total food and energy intake. This could include specific protein sources (e.g. beef, dairy or plant-derived [19]), certain amino acid combinations and hypothetically even micronutrients (e.g. [51]). Considering the role of specific amino acids in humans will become important as we come to identify the mechanistic bases for protein appetites. It seems increasingly likely that protein appetite is mediated by metabolic signalling that responds to circulating amino acids and the balance of specific amino acids within the dietary protein compartment [11,52,53]. If key amino acid combinations are identified to underpin protein appetites, versions of the analysis targeting these specific nutrients can be implemented. This might involve substituting the proportion of protein in the diet for another proportion entirely (e.g. proportion of non-branched chain amino acids [53]), or fitting the density of specific amino acids within the protein being consumed as a variable that moderates the effect of proportion of protein.

Nowhere is the nature of evidence more disputed than in the field of human nutrition, with diet surveillance data being especially contested [14,15,54–57]. However, to ignore these data altogether risks throwing the proverbial baby out with the bathwater. As discussed, a test of PL using cross-sectional data rests on the assumption that otherwise matched individuals have been subjected to risks throwing the proverbial baby out with the bathwater. As discussed, a test of PL using cross-sectional data rests on the assumption that otherwise matched individuals have been subjected to

Data accessibility. Code is maintained and publicly available from https://github.com/AlistairMcNairSenior/PL_Theory, and an archived version containing the exact analyses within the paper has been publicly released at https://zenodo.org/record/6643746#.Yqka5C8Rpqs (doi:10.5281/zenodo.6643746).

Authors’ contributions. A.M.S.: conceptualization, formal analysis, investigation, methodology, visualization, writing—original draft; D.R.: conceptualization, investigation, writing—review and editing; S.J.S.: conceptualization, investigation, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. This work was supported by a Program Grant from the Australian National Health and Medical Research Council (GNT1149976).

Acknowledgements. A.M.S. is supported by a J&D Coffey Fellowship from the University of Sydney, Charles Perkins Centre. S.J.S. and D.R. are supported by a Program Grant from the Australian National Health & Medical Research Council. All authors conceived the study and contributed to the manuscript and interpretation of results. A.M.S. performed the statistical modelling. The authors declare no conflicts of interest.
intake in the USA. Public Health Nutr. 21, 114–124. (doi:10.1017/s1368980017001574)

44. Monteiro CA et al. 2019 Ultra-processed foods: what they are and how to identify them. Public Health Nutr. 22, 936–941. (doi:10.1017/s1368980018003762)

45. Gibney MJ, Forde CG, Mullally D, Gibney ER. 2017 Ultra-processed foods in human health: a critical appraisal. Am. J. Clin. Nutr. 106, 717–724. (doi:10.3945/ajcn.117.160440)

46. Hall KD et al. 2019 Ultra-processed diets cause excess calorie intake and weight gain: an inpatient randomized controlled trial of ad libitum food intake. Cell Metab. 30, 67–77. e63. (doi:10.1016/j.cmet.2019.05.008).

47. Heitmann BL, Lissner L. 1995 Dietary underreporting by obese individuals—is it specific or non-specific? BMJ 311, 986. (doi:10.1136/bmj.311.7011.986)

48. Heerstrass D, Ocké M, Bueno-de-Mesquita H, Peeters P, Seidall J. 1990 Underreporting of energy, protein and potassium intake in relation to body mass index. Int. J. Epidemiol. 27, 186–193. (doi:10.1093/ije/27.2.186)

49. Goris AHC, Westerterp-Plantenga MS, Westerterp KR. 2000 Undereating and underreporting of habitual food intake in obese men: selective underreporting of fat intake. Am. J. Clin. Nutr. 71, 130–134. (doi:10.1093/ajcn/71.1.130)

50. Simpson SJ, Raubenheimer D. 2012 The nature of nutrition: a unifying framework. Aust. J. Zool. 59, 350–368.

51. Brunstrom JM, Schatzker M. 2022 Micronutrients and food choice: a case of ‘nutritional wisdom’ in humans? Appetite 174, 106055. (doi:10.1016/j.appet.2022.106055)

52. Solon-Biet SM, Griffiths L, Fosh S, Le Couteur DG, Simpson SJ, Senior AM. 2022 Meta-analysis links dietary branched-chain amino acids to metabolic health in rodents. BMC Biol. 20, 19. (doi:10.1186/s12915-021-01201-2)

53. Solon-Biet SM et al. 2019 Branched-chain amino acids impact health and lifespan indirectly via amino acid balance and appetite control. Nat. Metab. 1, 532–545. (doi:10.1038/s42255-019-0059-2)

54. Satija A, Stampfer MJ, Rimm EB, Willett W, Hu FB. 2018 Perspective: are large, simple trials the solution for nutrition research? Adv. Nutr. 9, 378–387. (doi:10.1093/advances/nmy030)

55. Mitka M. 2013 Do flawed data on caloric intake from NHANES present problems for researchers and policy makers? JAMA 310, 2137–2138. (doi:10.1001/jama.2013.281865)

56. Barnard ND, Willett WC, Ding EL. 2017 The misuse of meta-analysis in nutrition research. JAMA 318, 1435–1436. (doi:10.1001/jama.2017.12083)

57. Ioannidis JPA. 2018 The challenge of reforming nutritional epidemiologic research. JAMA 320, 969–970. (doi:10.1001/jama.2018.11025)