This page is intentionally blank. Proper e-companion title page, with INFORMS branding and exact metadata of the main paper, will be produced by the INFORMS office when the issue is being assembled.

E-companion to The Practical Value of Field Experiments

EC.1. Multiplicative Model

Consider an alternative multiplicative demand model of the following form:

$$\Delta q_i = x_1^{a_{i1}} x_2^{a_{i2}} \cdots x_n^{a_{in}} w_i.$$

Taking logs, we obtain

$$\begin{aligned}\log(\Delta q_i) &= a_{i1}\log(x_1) + a_{i2}\log(x_2) + \cdots + a_{in}\log(x_n) + \log(w_i) \\ &= \sum_{\ell=1}^{n} a_{i\ell}\log(x_\ell) + \log(w_i).\end{aligned}$$

By defining $\Delta\tilde{q}_i \triangleq \log(\Delta q_i), \tilde{x}_\ell \triangleq \log(x_\ell), \tilde{w}_i \triangleq \log(w_i)$, we can rewrite the above as

$$\Delta\tilde{q}_i = \sum_{\ell=1}^{n} a_{i\ell}\tilde{x}_\ell + \tilde{w}_i,$$

which is of the same form as our standard linear additive model.

Suppose that the noise term w_i is log-normally distributed and hence $\tilde{w}_i \sim N(0, c^2)$.[10] We are free to choose the decisions x_ℓ, and so let us choose each one randomly by first choosing u_ℓ uniformly from the interval $[-\rho, \rho]$ and then assigning $x_\ell = e^{u_\ell}$. Thus, $\tilde{x}_\ell \sim U[-\rho, \rho]$. We continue to assume independence among the x's and w's, which translates into independence among the $\tilde{x}$'s and $\tilde{w}$'s. Therefore, we can apply the same estimation method as described in Chapter 3 to learn the $\mathbf{A}$ matrix in this multiplicative model. In particular, the statistic defined in Section 3.5 becomes $\tilde{y}_{ij} \triangleq \beta \cdot \Delta\tilde{q}_i \cdot \tilde{x}_j$, which would again be an unbiased estimator of a_{ij}. In addition, our methodology for estimating k and d from empirical data and our simulation procedure, presented in Sections 4 and 5, can be similarly adapted to fit the multiplicative model.

EC.2. Asymptotic Notation

Let $\mathbf{n}$ be a vector of variables; then we say:

(i) $f(\mathbf{n}) \in O(g(\mathbf{n}))$ if there exist constants N and $C > 0$ such that $|f(\mathbf{n})| \leq C|g(\mathbf{n})|$ for all $\mathbf{n}$ such that $n_i > N,\ \forall i$;

(ii) $f(\mathbf{n}) \in \Omega(g(\mathbf{n}))$ if there exist constants N and $C > 0$ such that $|f(\mathbf{n})| \geq C|g(\mathbf{n})|$ for all $\mathbf{n}$ such that $n_i > N,\ \forall i$;

(iii) $f(\mathbf{n}) \in \Theta(g(\mathbf{n}))$ if $f(\mathbf{n}) \in O(g(\mathbf{n}))$ and $f(\mathbf{n}) \in \Omega(g(\mathbf{n}))$.

[10] More generally, we can relax this assumption – we require only that $\tilde{w}_i$ is sub-Gaussian with parameter c and has zero mean.

In the first case, $f(n) \in O(g(n))$ essentially means that $f(n)$ grows *no faster* than $g(n)$ as n becomes large. In this sense, $g(n)$ can be thought of as an "upper bound" on the rate of growth of $f(n)$. An example is $f(n) = 100n$ and $g(n) = n^2$.

In the second case, $f(n) \in \Omega(g(n))$ essentially means that $f(n)$ grows *at least as fast* as $g(n)$ as n becomes large. And so in this case, $g(n)$ can be thought of as a "lower bound" on the rate of growth of $f(n)$. An example is $f(n) = n$ and $g(n) = \log n + 100\sqrt{n}$.

In the last case, $f(n) \in \Theta(g(n))$ means that $f(n)$ and $g(n)$ grow at essentially the same rate as n becomes large. An example is $f(n) = n + \sqrt{n}$ and $g(n) = 2n - 1$, as both grow linearly with n. We say that $f(n) \in \Theta(n)$ and $g(n) \in \Theta(n)$.

As illustrated above, asymptotic notation focuses on the order of growth and ignores constants. To justify the importance of focusing on the order of growth in the regime of large numbers of products, let us consider the following example.

EXAMPLE EC.1. *(Impact of linear vs. logarithmic growth.)* Suppose that there are two estimation methods, requiring $s_1(n) = n$ and $s_2(n) = 10 \log n$ experiments, respectively, in order to estimate an $\mathbf{A}$ matrix for n products. For a small number of products, such as $n = 10$, the first method requires just 10 experiments, whereas the second method requires $10 \log(10) \approx 23$ experiments. However, with a large number of products, such as $n = 100$, the first method now requires 100 experiments, whereas the second method requires $10 \log(100) \approx 46$ experiments, a much smaller number. As the number of products increases further, the difference between the two methods becomes more and more pronounced.

The purpose of asymptotic notation is to focus on the dominant scaling factor and ignore constants, such as 10 in method 2 of the example above. Although these constants have a relatively larger impact when n is small, they become insignificant as n becomes large. Specifically, we say that for method 1, $s_1(n) \in \Theta(n)$, and for method 2, $s_2(n) \in \Theta(\log n)$.

EC.3. Proof of Theorem 1

THEOREM 1. (Estimation accuracy with sub-Gaussian noise for general $\mathbf{A}$ matrices.) *Under Assumption 1, for any $n \times n$ matrix $\mathbf{A}$ and any $\epsilon \geq 0$,*

$$\mathbf{P}\left(\max_{i,j} |\hat{a}_{ij} - a_{ij}| \geq \epsilon\right) \leq 2n^2 \exp\left\{-\frac{s\epsilon^2}{\max_i 36\left(\sum_{\ell=1}^{n} a_{i\ell}^2 + c^2/\rho^2\right)}\right\}.$$

Proof. Let our decisions be i.i.d. continuous random variables x distributed uniformly on $[-\rho, \rho]$, so that $\mathbb{E}[x] = 0$ and $\text{var}(x) = \mathbb{E}[x^2] = \rho^2/3$. We perform an experiment using a vector of decisions $\mathbf{x}$. Let Δq_i be the observed percentage change in demand for product i, and let x_j be the pricing decision for product j.

Having defined $\beta \triangleq 3/\rho^2$, we consider the statistic

$$y_{ij} = \beta(\Delta q_i x_j) = \beta\left(\sum_{\ell=1}^{n} a_{i\ell}x_\ell + w_i\right)x_j,$$

which satisfies $\mathbb{E}[y_{ij}] = a_{ij}$. Therefore, y_{ij} is an unbiased estimator of a_{ij}. Let $y_{ij}(t)$ be the statistic calculated from the t^{th} experiment. By Assumption 1, for each (i,j), the statistics $y_{ij}(t)$ are independent and identically distributed across different experiments t. By the law of large numbers, the sample mean $\hat{a}_{ij} \triangleq \frac{1}{s}\sum_{t=1}^{s} y_{ij}(t)$ converges to a_{ij} as we take many samples from many experiments. We wish to bound the concentration of $\hat{a}_{ij}$ around its mean, a_{ij}.

To do so, we show that $\hat{a}_{ij}$ is sub-Gaussian. A random variable X is sub-Gaussian with parameter $\sigma > 0$ if

$$\mathbb{E}[\exp(\lambda(X - \mathbb{E}[X]))] \le \exp(\sigma^2\lambda^2/2) \tag{EC.1}$$

for all $\lambda \in \mathbb{R}$. We make use of the following well-known properties:

1. If X is sub-Gaussian with parameter σ, then $aX + b$ is sub-Gaussian with parameter $|a|\sigma$.

2. If X is bounded a.s. in an interval $[a,b]$, then X is sub-Gaussian with parameter at most $(b-a)/2$.

3. If X_1 and X_2 are sub-Gaussian with parameters σ_1 and σ_2, respectively,

   (a) and if X_1 and X_2 are independent, then $X_1 + X_2$ is sub-Gaussian with parameter $\sqrt{\sigma_1^2 + \sigma_2^2}$.

   (b) and if X_1 and X_2 are not independent, then $X_1 + X_2$ is sub-Gaussian with parameter at most $\sqrt{2(\sigma_1^2 + \sigma_2^2)}$.

4. If X is sub-Gaussian with parameter σ, then it satisfies the following concentration bound:

$$\mathbf{P}(|X - \mathbb{E}[X]| \ge \epsilon) \le 2\exp\left(-\frac{\epsilon^2}{2\sigma^2}\right), \quad \forall \epsilon \ge 0. \tag{EC.2}$$

We first consider the random variable y_{ij}:

$$\begin{aligned} y_{ij} &= \beta\left(\sum_{\ell=1}^{n} a_{i\ell}x_\ell + w_i\right)x_j \\ &= \beta\left\{\left(\sum_{\ell\neq j} a_{i\ell}x_\ell + w_i\right)x_j + a_{ij}x_j^2\right\} \\ &= \beta\left\{Vx_j + a_{ij}x_j^2\right\}, \end{aligned}$$

where we have defined

$$V \triangleq \sum_{\ell\neq j} a_{i\ell}x_\ell + w_i.$$

We now show that V is sub-Gaussian. For each ℓ, x_ℓ is bounded on $[-\rho,\rho]$ and therefore sub-Gaussian with parameter ρ. Hence, $a_{i\ell}x_\ell$ is sub-Gaussian with parameter $|a_{i\ell}|\rho$. Also, under

Assumption 1, w_i is sub-Gaussian with parameter c. The random variables $a_{i\ell}x_\ell$ and w_i are all independent. Therefore, their sum, V, is also sub-Gaussian with parameter $\sigma_V \triangleq \sqrt{\sum_{\ell\neq j} a_{i\ell}^2\rho^2 + c^2}$.

Next, we show that Vx_j is sub-Gaussian using the definition. For any $\lambda \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{E}\left[\exp\left\{\lambda(Vx_j - \mathbb{E}[Vx_j])\right\}\right] &= \mathbb{E}\left[\exp\left\{\lambda(Vx_j)\right\}\right] && \text{(EC.3)}\\
&= \int_{-\rho}^{\rho} \mathbb{E}[\exp\{\lambda(Vx)\}]\frac{1}{2\rho}\,dx && \text{(EC.4)}\\
&\leq \int_{-\rho}^{\rho} \exp\left\{(|x|\sigma_V)^2\lambda^2/2\right\}\frac{1}{2\rho}\,dx && \text{(EC.5)}\\
&\leq \int_{-\rho}^{\rho} \exp\left\{(\rho\sigma_V)^2\lambda^2/2\right\}\frac{1}{2\rho}\,dx\\
&= \exp\left\{(\rho\sigma_V)^2\lambda^2/2\right\},
\end{aligned}$$

where (EC.3) is because Vx_j has zero mean; (EC.4) is obtained by conditioning on the values of x_j; and (EC.5) follows from (EC.1) and the fact that for any $x \in [-\rho, \rho]$, Vx is zero-mean and sub-Gaussian with parameter $|x|\sigma_V$. Therefore, Vx_j is also sub-Gaussian with parameter $\rho\sigma_V$.

Next, we show that $a_{ij}x_j^2$ is sub-Gaussian. Since x_j^2 is bounded in $[0, \rho^2]$, it is sub-Gaussian with parameter $\rho^2/2$. Therefore, $a_{ij}x_j^2$ is sub-Gaussian with parameter $\rho^2|a_{ij}|/2$.

Finally, y_{ij} is a sum of two (dependent) sub-Gaussian random variables: βVx_j with parameter $\beta\rho\sigma_V$, and $\beta a_{ij}x_j^2$ with parameter $\beta\rho^2|a_{ij}|/2$. Therefore, y_{ij} is also sub-Gaussian with parameter

$$\begin{aligned}
\sigma_Y \triangleq \sqrt{2(\beta^2\rho^2\sigma_V^2 + \beta^2\rho^4 a_{ij}^2/4)} &= \sqrt{2\left\{\beta^2\rho^2\left(\sum_{\ell\neq j} a_{i\ell}^2\rho^2 + c^2\right) + \beta^2\rho^4 a_{ij}^2/4\right\}}\\
&\leq \sqrt{2\beta^2\rho^4\left(\sum_{\ell=1}^{n} a_{i\ell}^2 + c^2/\rho^2\right)}\\
&= \sqrt{18\left(\sum_{\ell=1}^{n} a_{i\ell}^2 + c^2/\rho^2\right)}.
\end{aligned}$$

Since $\hat{a}_{ij} = \frac{1}{s}\sum_{t=1}^{s} y_{ij}(t)$ is a sample mean of s independent y_{ij}'s, $\hat{a}_{ij}$ is sub-Gaussian with parameter

$$\sigma_{ij} \triangleq \frac{1}{s}\sqrt{s\sigma_Y^2} \leq \sqrt{\frac{18}{s}\cdot\left(\sum_{\ell=1}^{n} a_{i\ell}^2 + c^2/\rho^2\right)}.$$

We can then bound the concentration of our estimator $\hat{a}_{ij}$ around the true parameter a_{ij} using (EC.2):

$$\begin{aligned}
\mathbf{P}(|\hat{a}_{ij} - a_{ij}| \geq \epsilon) &\leq 2\exp\left\{-\frac{\epsilon^2}{2\sigma_{ij}^2}\right\}\\
&\leq 2\exp\left\{-\frac{s\epsilon^2}{36\left(\sum_{\ell=1}^{n} a_{i\ell}^2 + c^2/\rho^2\right)}\right\}.
\end{aligned}$$

This gives a concentration bound for the error of a particular (i,j) pair. To arrive at the final result, which bounds the maximum error over all (i,j) pairs, we apply the union bound and conclude that

$$\mathbf{P}\left(\max_{i,j}|\hat{a}_{ij}-a_{ij}|\geq\epsilon\right)\leq 2n^2\exp\left\{-\frac{s\epsilon^2}{\max_i 36\left(\sum_{\ell=1}^n a_{i\ell}^2+c^2/\rho^2\right)}\right\}. \quad \square$$

EC.4. Proof of Theorem 2

THEOREM 2. (Necessary condition for uniform ϵ-accurate estimation under sparsity with Gaussian noise.) *For $\lambda>0$, let*

$$\mathcal{A}_{n,k}(\lambda)\triangleq\left\{\mathbf{A}\in\mathbb{R}^{n\times n}:|\{j:a_{ij}\neq 0\}|=k,\forall i=1,\ldots,n;\min_{i,j:a_{ij}\neq 0}|a_{ij}|\geq\lambda\right\}$$

be the class of $n\times n$ $\mathbf{A}$ matrices whose rows are k-sparse and whose nonzero entries are at least λ in magnitude. Let the noise terms be i.i.d. $\mathcal{N}(0,c^2)$ for some $c>0$. Suppose that for some $\epsilon\in(0,\lambda/2)$ and $\delta\in(0,1/2)$, we have an estimator that

(a) experiments with percentage price changes $x\in[-1,\tilde{\rho}]$, for some $\tilde{\rho}\geq 1$ (i.e., the price of each product cannot fall below 0 and cannot increase by more than $100\tilde{\rho}\%$), and

(b) for any $\mathbf{A}$ matrix in $\mathcal{A}_{n,k}(\lambda)$ achieves uniformly ϵ-accurate estimates with probability $1-\delta$. Then, the number of experiments used by the estimator must be at least

$$s\geq\frac{k\log(n/k)-2}{\log(1+k^2\lambda^2\tilde{\rho}^2/c^2)}.$$

Proof. For any $\lambda>0$, consider the class $\mathcal{A}_{n,k}(\lambda)$. Fix some $\epsilon\in(0,\lambda/2)$ and $\delta\in(0,1/2)$. In what follows, all estimators use the results of s experiments, for some arbitrary s.

Define the sub-class $\mathcal{A}_{n,k}^{\text{const}}(\lambda)\triangleq\{\mathbf{A}\in\mathbb{R}^{n\times n}:|\{j:a_{ij}\neq 0\}|=k,\forall i=1,\ldots,n;a_{ij}=\lambda,\forall i,j \text{ s.t. } a_{ij}\neq 0\}\subset\mathcal{A}_{n,k}(\lambda)$, which is the class of all $n\times n$ $\mathbf{A}$ matrices whose rows are k-sparse and whose nonzero entries are all exactly equal to λ.

The desired specification is an estimator that for any $\mathbf{A}$ matrix in $\mathcal{A}_{n,k}(\lambda)$ achieves uniformly ϵ-accurate estimates with probability $1-\delta$. In order to obtain a lower bound on the number of experiments needed to meet this specification, it suffices to obtain a lower bound on the number of experiments needed to meet the following looser specification: we let the $\mathbf{A}$ matrix be generated uniformly at random from the sub-class $\mathcal{A}_{n,k}^{\text{const}}(\lambda)$ and require that with probability at least $1-\delta$ the first row of $\mathbf{A}$ is correctly recovered to uniform ϵ-accuracy. Because $\mathbf{A}\in\mathcal{A}_{n,k}^{\text{const}}(\lambda)$, all elements of $\mathbf{A}$ are either exactly 0 or λ, and since $\epsilon\in(0,\lambda/2)$, achieving uniform ϵ-accuracy is equivalent to perfectly recovering $\mathbf{A}$, which is also equivalent to perfectly recovering the sparsity pattern of $\mathbf{A}$ (i.e., identifying the locations of all nonzero entries). Let R_1^{const} denote the event of exactly recovering the sparsity pattern of the first row of an $\mathbf{A}$ matrix chosen uniformly at random from $\mathcal{A}_{n,k}^{\text{const}}(\lambda)$.

We now focus on the event R_1^{const} and find an upper bound on its probability. Within the sub-class $\mathcal{A}_{n,k}^{\text{const}}(\lambda)$, there are exactly $N \triangleq \binom{n}{k}$ possible sparsity patterns for the first row of any $\mathbf{A}$ matrix. Moreover, because all nonzero entries are equal to the same value λ, each unique sparsity pattern corresponds to a unique row vector, and vice versa. Suppose that we randomly choose the first row $\mathbf{a}_1'$ by choosing one of the N possible sparsity patterns uniformly at random. We can then view the sparsity pattern recovery problem as a channel coding problem. The randomly selected sparsity pattern $\theta \in \{1, \ldots, N\}$ is encoded, using a sequence of s experimental decisions $\mathbf{X} \in \mathbb{R}^{n \times s}$, into codewords $\mathbf{r} = \mathbf{a}_1'\mathbf{X} = (r_1, r_2, \ldots, r_s) \in \mathbb{R}^s$. These codewords represent the uncorrupted percentage change in demand for product 1 in each of the s experiments. The codewords are sent over a Gaussian channel subject to noise $\mathbf{w} = (w_1, w_2, \ldots, w_s) \sim \mathcal{N}(0, c^2 I) \in \mathbb{R}^s$ and received as noisy measurements $\mathbf{y} = \mathbf{r} + \mathbf{w} = (y_1, y_2, \ldots, y_s) \in \mathbb{R}^s$, which are equal to the observed noisy percentage change in demand, $\Delta\mathbf{q}_1$. The goal is to recover the pattern θ from the measurements $\mathbf{y}$.

The power of a Gaussian channel is given by $P = \frac{1}{s}\sum_{t=1}^{s} r_t^2$. Since $\mathbf{a}_1'$ is k-sparse and any decision x is bounded in $[-1, \tilde{\rho}]$, we have that $|r_t| \le k\lambda\tilde{\rho}$ for all t, and hence $P \le k^2\lambda^2\tilde{\rho}^2$. From standard results (Cover and Thomas 1991), the capacity of a Gaussian channel with power P and noise variance c^2 is $\frac{1}{2}\log\left(1 + \frac{P}{c^2}\right)$. Therefore, the capacity of our particular channel is

$$C \le \frac{1}{2}\log\left(1 + \frac{k^2\lambda^2\tilde{\rho}^2}{c^2}\right).$$

From Fano's inequality (Cover and Thomas 1991), we know that the probability of error, P_e, of a decoder that decodes the sparsity pattern θ from noisy measurements $\mathbf{y} \in \mathbb{R}^s$ is lower bounded as

$$\begin{aligned} P_e &\ge \frac{H(\theta \mid \mathbf{y}) - 1}{\log N} \\ &= \frac{H(\theta) - I(\theta; \mathbf{y}) - 1}{\log N} \\ &= \frac{\log N - I(\theta; \mathbf{y}) - 1}{\log N} \\ &= 1 - \frac{I(\theta; \mathbf{y}) + 1}{\log N}, \end{aligned}$$

where H denotes entropy and I denotes mutual information. The first equality is by the definition of mutual information, and the second equality follows from the fact that θ is chosen uniformly over a set of cardinality N. We can upper bound the mutual information between θ and $\mathbf{y}$ as

$$\begin{aligned} I(\theta; \mathbf{y}) &\le I(\mathbf{r}; \mathbf{y}) \\ &= h(\mathbf{y}) - h(\mathbf{y} \mid \mathbf{r}) \\ &= h(\mathbf{y}) - h(\mathbf{w}) \end{aligned} \tag{EC.6}$$

$$\leq \sum_{t=1}^{s} h(y_t) - \sum_{t=1}^{s} h(w_t) \tag{EC.7}$$

$$= \sum_{t=1}^{s} [h(y_t) - h(y_t \mid r_t)]$$

$$= \sum_{t=1}^{s} I(r_t; y_t)$$

$$\leq sC, \tag{EC.8}$$

where h denotes differential entropy, (EC.6) follows from the data processing inequality, (EC.7) follows from the independence of the w_t's and the fact that the entropy of a collection of random variables $\{y_t\}$ is no more than the sum of their individual entropies, and (EC.8) follows from the definition of channel capacity as the maximal mutual information. And so by Fano's inequality, the probability of error is lower bounded by

$$P_e \geq 1 - \frac{sC+1}{\log N},$$

which immediately gives the following upper bound on the probability of R_1^{const}:

$$\mathbf{P}(R_1^{\text{const}}) = 1 - P_e \leq \frac{sC+1}{\log N}.$$

Therefore, achieving the looser specification of uniform ϵ-accurate estimates of the first row of a random $\mathbf{A} \in \mathcal{A}_{n,k}^{\text{const}}(\lambda)$ with probability $1-\delta$ implies the following condition on the number of experiments, s:

$$1 - \delta \leq \frac{sC+1}{\log N} \Longrightarrow s \geq \frac{(1-\delta)\log N - 1}{C}.$$

Consequently, achieving the stricter original specification of an estimator that for all $\mathbf{A}$ matrices in $\mathcal{A}_{n,k}(\lambda)$ achieves uniformly ϵ-accurate estimates with probability $1-\delta$ also requires the number of experiments to satisfy the above condition.

With some simple rearrangement, and noting that $\log N = \log \binom{n}{k} \geq k \log(n/k)$ and $\delta \in (0, 1/2)$, we obtain the desired lower bound:

$$s \geq \frac{(1-\delta)\log N - 1}{C} \geq \frac{2(1-\delta)k\log(n/k) - 2}{\log(1 + k^2\lambda^2\tilde{\rho}^2/c^2)} \geq \frac{k\log(n/k) - 2}{\log(1 + k^2\lambda^2\tilde{\rho}^2/c^2)}. \quad \square$$