# Replication of Berry et al. (1995)

Matthew Gentzkow*
*Stanford and NBER*
Jesse M. Shapiro
*Brown and NBER*

October 2016

## Replication Code and Data

## A Archive Overview

Directories within `/analysis/` produce tables reported in the replication document, and directories within `/derived/` produce derived data used in `/analysis/`. Table 1 summarizes the contents of each directory.

In relevant directories, the subdirectory `/external/lib/matlab/blp/` contains our MATLAB implementation of BLP (1995), and the subdirectory `/external/lib/blp_1999/` contains data and GAUSS code for BLP (1999).[1]

## B Directory Structure

Each directory is self-contained in the sense that the code does not require inputs from outside the directory. Each `/derived/` and `/analysis/` directory follows a similar structure:

- `/code/` - Contains all scripts necessary to execute directory.
- `/code/external.txt` - List of external resources needed. These are prepopulated in subdirectory `/external/`.
- `/code/make.py` - Code to execute scripts in the required order. You will not be able to run this code. Treat it as documentation on the correct order of the other scripts.
- `/external/` - Contains data and code modules necessary to execute the directory.
- `/output/` - Contains LOG files, TXT files with tables, and MATLAB data files.

*E-mail: gentzkow@stanford.edu, jesse_shapiro_1@brown.edu.

[1] Accessed on July 16, 2014 via Archive.org's cached version: https://web.archive.org/web/20050430123542/http://www-personal.umich.edu/~jamesl/verstuff/instructions.html

# C System Requirements

The directories were originally run on a 64-bit Windows 7 machine with Stata/MP 13 and Matlab 2012b.

# D Notes

All notation is defined in BLP (1995) unless explicitly defined below.

## D.1 Raw Inputs

**Firm/Model data:** Using the BLP (1999) replication code and data, we are able to extract the following variables: car id, model id, firm id, year, market share, quantity, price, indicators for production location, horsepower per weight, whether air conditioning comes standard, miles per dollar, size of car, miles per gallon, and the logit dependent variable $(\ln(s_j) - \ln(s_0))$.[2] All other firm/model variables used are transformations of these base variables.

**Unobservables:** For each individual in the model, we draw a vector of $v_i = (v_{iy}, v_{i1} \ldots, v_{iK})$ unobservables from a standard multivariate normal distribution with identity variance-covariance matrix, where we set $y_i = e^{m_t + \hat{\sigma}_y v_{iy}}$. Note that the unobservables are fixed over the time period of the panel (see BLP (1995) p. 868). Define $n_s = 200$ to be the number of unobservables used in simulation.[3] Importance sampling is done directly in BLP (1999) code, which gives a corresponding importance sampling weight $w_i^u$ for each unobservable $i$.[4] We use the weight when taking integrals across unobservables. The BLP (1999) code also provides the mean of income for each year $m_t$ and standard deviation of income $\sigma_y$ across all years in the sample.

## D.2 Deviations and Clarifications from BLP (1995)

**Utility:** Following BLP (1999) code, we can approximate utility with

$$u_{itj} = -\frac{\alpha p_{tj}}{e^{m_t + \hat{\sigma}_y v_{iy}}} + x_{tj}\overline{\beta} + \xi_{tj} + \sum_k \sigma_k x_{tjk} v_{ik} + \varepsilon_{itj},$$

which implies the outside good's utility is normalized to 0.

**Delta contraction mapping:** We use the logit estimate (estimate with no random coefficients) as the starting

[2] We redefine a car to be from Japan if it is not produced in Europe and not produced domestically in the US. BLP (1999) makes additional restrictions, but our definition matches BLP (1995)'s summary statistics. Also, BLP (1995) mention that they use 997 models (see p. 869), but BLP (1999) code returns 999 models instead.

[3] Note that this number matches the discussion in BLP's NBER Working Paper No. 4264, footnote 17.

[4] Note that importance sampling requires computation of market shares and thus relies on a set of parameter inputs. By default, BLP (1999) code uses a set of parameter guesses, which it also uses as the starting values for estimation. Since we ultimately decided to use BLP (1995) published parameters as our starting values, we have altered BLP (1999) code so that it uses the published parameters for importance sampling instead. Thus, our unobservable draws differ slightly from the default ones from BLP (1999) code. For our bootstrap procedure, we draw 10 sets of 10 unobservables for each year following the same procedure. Then, for each block bootstrap at the market level, we use the first set of unobservables for the first time a market year is selected, the second set of unobservables for the second time a market year is selected, etc. This gives a sample of 200 unobservables that are fixed across the bootstrap panel. We repeat this for each bootstrap iteration.

value for every iteration of the contraction mapping. For each market $t$, we continue until

$$\max_j \left( \left| \frac{\delta_{tj,\text{prev}}}{\delta_{tj,\text{new}}} - 1 \right| \right) \leq 10^{-14}.$$

**Market shares:** When we compute individual market shares, we treat each market $t$ separately. Also, when we integrate over the distribution of $v_i$ to obtain the market shares conditional only on product attributes, we use:

$$s_j(p,x,\xi,\theta,P_{n_s}) = \frac{1}{n_s}\sum_{i=1}^{n_s} f_j(v_i,\delta,p,x,\theta)\, w_i^u,$$

where $f_j(v_i,\delta,p,x,\theta)$ are $j$'s market shares conditional on $v_i$.

**Marginal costs:** We simulate the integrals defined in BLP (1995) equations (6.9a) and (6.9b) in a manner similar to our simulations for market shares. For example, (6.9a) is

$$\frac{1}{n_s}\sum_{i=1}^{n_s} f_j(v_i,\delta,x,p,\theta)\,(1 - f_j(v_i,\delta,x,p,\theta))\,[\partial\mu_{ij}/\partial p_j]\, w_i^u,$$

where $\partial\mu_{ij}/\partial p_j = \frac{-\alpha}{e^{m_t + \sigma_y v_{iy}}}$. Note that the derivative of market share with respect to price is 0 *across* markets, even within the same firm.

**Instruments:** The set of instruments defined in BLP (1995)[5] is likely to be highly collinear. To solve this problem, we remove collinearity as follows:

1. Start with the first instrument, let this be $x$.
2. Choose the next instrument, let this be $y$.
3. Regress $y$ onto $x$. If the $R^2$ is lower than the tolerance, accept $y$ as a new instrument. If the $R^2$ is higher than the tolerance, drop $y$.
4. Redefine $y$ to be the next instrument, and redefine $x$ to be all accepted instruments.
5. Repeat steps 4-6 until all instruments have either been accepted or dropped.

Following BLP (1999) code, we choose an $R^2$ tolerance of 0.99 and do this collinearity check separately for the demand- and supply-side instruments.[6] After removing collinear instruments, we demean all remaining, non-constant instruments prior to estimation to reduce mechanical correlation among the moments.

**IV Regression:** We combine the supply-side and demand-side IV regressions together into a single regression, which corresponds to the actual procedure done in the BLP (1999) code. Let $x$ ($z_1$) denote the

[5] Based on BLP (1999) code, it appears as if BLP (1995) had a mistake in the implementation of their instruments. When creating 'own' firm instruments, instead of summing across a firm's other models, they multiplied the instrument by the number of models the firm was producing. Replication of the IV logit specification in table 3 of BLP requires using these alternative instruments. Our code replicates the original coefficients and we have included comments to indicate where we have deviated from the definitions in the text of the article.

[6] Consider $z_s^{own}$ = {constant, horsepower per weight, air conditioning, miles per dollar, and size} and $z_d^{own}$ = {constant, log of horsepower per weight, air conditioning, log of mpg, log of size, and year trends}. For the supply-side instruments, the order is: each element of $z_s^{own}$ in order, own firm instruments in same order, rival firm instruments in same order, and own miles per dollar. For the demand-side instruments, the order is: each element of $z_d^{own}$ in order, own firm instruments in same order, and rival firm instruments in same order.

demand-side variables (instruments) and $w$ ($z_2$) denote the supply-side variables (instruments). Denote

$$X = \begin{pmatrix} x & 0 \\ 0 & w \end{pmatrix}$$

and define $Z$ similarly. Perform 2SLS on

$$\begin{bmatrix} \delta \\ \log(mc) \end{bmatrix} = X \begin{bmatrix} \beta \\ \gamma \end{bmatrix} + \begin{bmatrix} \xi \\ \omega \end{bmatrix}.$$

The 2SLS projection matrix used is $ZWZ'$, where $W$ is the appropriate GMM weight matrix from below, unless explicitely stated otherwise. Using the 2SLS coefficient estimates $(\hat{\beta}, \hat{\gamma})'$, one can back-out estimates of the unobservables $(\hat{\xi}, \hat{\omega})'$.

**Moments:** Similar to BLP (1995) p. 863, the GMM distance vector $g$ is defined as the average across models $m$ of

$$\hat{g}_m(\theta) = \begin{bmatrix} \sum_t Z'_{dmt} \otimes \hat{\xi}_i(\theta) \\ \sum_t Z'_{smt} \otimes \hat{\omega}_i(\theta) \end{bmatrix}$$

where $Z$ is as defined above.

## D.3 Estimation Steps

As in BLP (1999) code, we use two-stage GMM to estimate the model. In summary, the steps are:

1. Determine a set of starting random coefficient parameters, $(\alpha, \sigma)_{\text{start}}$. We use BLP (1995) published values as our starting parameters.
2. Using the standard 2SLS projection matrix for IV regression $Z(Z'Z)^{-1}Z'$, compute the distance vector $\hat{g}_m$ for *each* model $m$ at $(\alpha, \sigma)_{\text{start}}$.
3. Define the initial GMM weight matrix $W_{\text{GMM1}}$ as the inverse of the variance-covariance matrix for the distance vectors $\hat{g}_m$.
4. Using $(\alpha, \sigma)_{\text{start}}$ and $W_{\text{GMM1}}$, estimate the first-stage parameter estimates $(\alpha, \sigma)_{\text{GMM1}}$ via GMM.
5. Compute the new weight matrix $W_{\text{GMM2}}$ in the same manner as before, but at $(\alpha, \sigma)_{\text{GMM1}}$ and using $W_{\text{GMM1}}$ for the 2SLS weight matrix instead.
6. Using $(\alpha, \sigma)_{\text{GMM1}}$ and $W_{\text{GMM2}}$, estimate the final parameter estimates $(\alpha, \sigma)_{\text{GMM2}}$ via GMM. Extract $\beta$ and $\gamma$ using $(\alpha, \sigma)_{\text{GMM2}}$.

We constrain all elements of $\sigma$ to be non-negative. We use Artelys Knitro's Interior/Direct algorithm as our solver.

## D.4 Sensitivity

The Jacobian $\hat{G}$ of the distance vectors with respect to the parameters is computed numerically using small perturbations to the estimated parameters. When computing the Jacobian for the random coefficient parameters $(\alpha, \sigma)$, the IV regression is re-estimated to obtain new error terms. When computing the Jacobian for the mean parameters $(\beta,\gamma)$, the IV regression is skipped and the error terms are instead computed with the perturbed mean parameters. For the weight matrix in the sensitivity calculation, we use $\hat{W}_g = W_{\text{GMM2}}$ from above. The moment variance-covariance matrix $\hat{\Omega}_{gg}$ is just the sample variance-covariance matrix of $\hat{g}(\theta)$ taken across models. We then compute the parameter variance-covariance matrix $\Sigma_{\theta\theta}$ using theorem (13.1) in Greene (2012), where $\theta$ contains both the random coefficient paramters $(\alpha,\sigma)$ and the mean parameters $(\beta,\gamma)$. Sensitivity and standardized sensitivity are then computed using these objects. Our sample sensitivity and standarized sample sensitivity computations use the matrix $\hat{A}$ in addition to these objects. When computing the $p$th column of $\hat{A}$ as $\hat{G}_p'\hat{W}_g\hat{g}_m$, we determine each matrix $\hat{G}_p$ numerically by calculating the change in $\hat{G}$, the Jacobian as computed using the method described above, associated with small perturbations to the estimated parameters. We use the number of models $n_m$, instead of the number of cars $n$ for the number of observations in all post-estimation objects.

**Table 1:** Archive overview

| **Directory name** | **Purpose** |
|---|---|
| /derived/Transparent Identification (BLP Extract Data)/ | Prepares and extracts BLP (1999) data |
| /derived/Transparent Identification (BLP Extract Unobservables)/ | Extracts unobservables for BLP (1995) model |
| /derived/Transparent Identification (BLP Verify Elasticity)/ | Outputs elasticities from GAUSS for MATLAB unit tests |
| /derived/Transparent Identification (BLP Verify Objective)/ | Outputs objective function information from GAUSS for MATLAB unit tests |
| /analysis/Transparent Identification (BLP Estimation)/ | Estimates BLP (1995) model |
| /analysis/Transparent Identification (BLP Replication)/ | Estimates descriptive results, computes elasticities and markups, and outputs all tables |