Readme File for "Copyright and Generic Entry in Book Publishing"

Three datasets are used in this paper:

The first, "amazon_data.dta", provides information for each *collected* edition on Amazon and Project Gutenberg (if available) on a monthly level (from April 2011 to April 2012), including information on prices, availability, demand, and other Amazon-specific characteristics, as well as information on the title-level, including the original year of publication, prizes won, British library checkouts, and membership in Harry Bloom's Western Canon.

A summary of the dataset is here:

```
Contains data from amazon_data.dta
  obs:        10,732
 vars:            31                          3 Oct 2018 20:33
 size:     3,766,932
```

variable name	storage type	display format	value label	variable label
title	str37	%37s		Title
author	str38	%38s		Author
format	str9	%9s		Edition format (hardcover, paperback, e-book)
asin	str10	%10s		Unique Amazon identifier
isbn	str14	%14s		Standard edition identifier
publisher	str196	%196s		Name of publisher (this edition)
year	int	%9.0g		Original year of publication (title)
post1923	byte	%9.0g		= 1 if the title is still protected
ntitle	int	%9.0g		# in-print editions per title
ntitleformat	byte	%9.0g		# in-print editions per title-format
t	byte	%9.0g		Month identifier
sales	int	%9.0g		Monthly sales (this edition)
gb	int	%9.0g		Monthly Gutenberg download count (title)
demand	int	%9.0g		= sales + gb
plr	int	%9.0g		British library checkouts
amazonprice	float	%9.0g		Price on Amazon (this edition)
newprice	float	%9.0g		Cheapest new price on Amazon (this edition)
usedprice	float	%9.0g		Cheapest used price on Amazon (this edition)
price	float	%9.0g		Cheapest avail. new price, or used price if no new
newnumb	int	%9.0g		# of sellers offering edition as new
usednumb	int	%9.0g		# of sellers offering edition as used
prize	byte	%9.0g		= 1 if the book received a Pulitzer Prize
prize_author	byte	%9.0g		= 1 if the author received a Pulitzer Prize
canon_title	byte	%9.0g		= 1 if title is listed in Western Canon
canon_author	byte	%9.0g		= 1 if author is listed in Western Canon
pages	int	%8.0g		# of pages (this edition)
picture	byte	%9.0g		= 1 if the Amazon page has a picture
age	int	%9.0g		Age in months (this edition)
major	byte	%9.0g		= 1 if published by a major publisher (this edition)
new	byte	%9.0g		= 1 if a new edition is available
used	byte	%9.0g		= 1 if a used edition is available

The second dataset, "edition_details.dta," reports information on *all* published editions for each title, as obtained from the Bowker Books-in-Print directory. This dataset is used to provide supporting evidence that the identifying assumptions of the regression discontinuity design are satisfied (see lines 187 – 216 of rdd_analysis.do). A description of the variables in this dataset is below:

```
Contains data from edition_details.dta
  obs:         4,930
 vars:            11                          3 Oct 2018 20:33
 size:       793,730
```

variable name	storage type	display format	value label	variable label
title	str37	%37s		Title
year	int	%9.0g		Original year of publication (title)
format	str9	%9s		Edition format (hardcover, paperback, e-book, etc)
isbn13	str17	%17s		Unique edition identifier
bowkerpublisher	str61	%61s		Publisher (this edition)
pubyear	int	%9.0g		Year of publication (this edition)
pubmonth	byte	%9.0g		Month of publication (this edition)
pubday	byte	%9.0g		Day of publication (this edition)
itemstatus	str25	%25s		In-print status of the edition
bowkerprice	float	%9.0g		Suggested retail price as listed on Bowker
pages	int	%9.0g		Number of pages (this edition)

The third dataset, "Gutenberg_downloads.dta" provides Project Gutenberg download counts over a 30-day period (April 2014) on 29,290 documents that are available on Project Gutenberg. See http://gutenberg.readingroo.ms/cache/generated/feeds/. This file is used for scaling up the calculated consumer surplus changes to an industry level.

<u>These datasets are used in two do-files:</u>

The first, "rdd_analysis_full.do," provides code for summary statistics (lines 9-30) and for all regressions in the regression discontinuity design, including main tables and robustness check (lines 34-147) and figures and evidence for identification (lines 152-216).

This do-file uses "amazon_data.dta" to estimate the counts and prices of editions. It uses "edition_details.dta" to provide evidence that the identification assumptions are satisfied.

The second do-file, "demand_cons_surplus.do," provides code to estimate demand for book editions on Amazon (lines 29-135), including robustness checks (lines 140-226). The do-file then creates figures to illustrate title qualities (lines 231-285). It then calculates consumer surplus based on the estimated demand function, per title (lines 289-409) and aggregated to the industry level (lines 414-478). Finally, it calculates marginal costs (lines 482-537), market shares (lines 541-564), annual operating profits and fixed costs (assuming that profits are zero under free entry in the public domain) (lines 568-689), and implied title profits for copyright-protected titles (lines 693-701).

This do-file uses "amazon_data.dta" for all analyses. It supplements these data with "Gutenberg_downloads.dta" to scale consumer surplus to the industry-level.