Product review, an insight about consumer search

Meisam Hejazinia

March 23, 2013

Outline

Data and Definitions

- Data
  - Firefox Download them all plug in product reviews for one year (365 days)
  - Number of of product reviews availabe: 1212
  - Number of categories available 20
  - Variation of downloads between 12,000 to 26,000 per day
  - Stem of words used by reviewers: 2603 stems categorized by human reviwer
- Notation:
  - Ψ Word category matrix $[2603 \times 20]$
  - Θ Review word matrix $[1212 \times 2603]$
  - $\Omega = \Theta * \Psi$ Review category matrix $[1212 \times 20]$

Data and Definitions cont.

- List of categories:
  - Information source, complement products, alternative products, experience, lack of experience
  - usage, product attribute, product human interaction, selection criteria
  - Cost (time and effort), Terminology, positive valance, negative valance, comparison
  - Social network, feeling intensity, uncertainty, politeness, broad, narrow

Model

- Model 1: Effect of review attributes on download count
  - $D = \beta * \Omega + \Sigma$
  - where D is number of daily download vector
- Variation of Model 1
  - Frequentist regression
  - poisson regression
  - Bayesian regression
  - Tobit regression
- Model 2: Effect of review attributes on number of star selection
  - $S = \beta * \Phi + \Sigma$ where:
  - S is number of stars selected
  - Φ is attribute of the review (in term of categories)
  - Σ would have normal error term
  - Model would be Ordinal Probit

Model cont.

- Model 3: Effect of review attributes on selection of each star level
  - $S_i = \beta * \Phi + \Sigma$ where:
  - S_i is whether i star is selected
  - Φ is attribute of the review (in term of categories)
  - Σ would have be normal or Gambel error term
  - Model: both binary probit and binary logit is checked

Variable	Estimate	Std. Error	$Pr(>\|t\|)$
(Interc.)	1.939e+04	2.040e+02	¡ 2e-16 ***
Info Src	1.215e+04	3.781e+03	0.001436 **
complement	2.041e+03	4.664e+02	1.61e-05 ***
substitute	-3.915e+03	1.164e+03	0.000855 ***
experience	1.784e+03	1.207e+03	0.140302
Naive	-2.602e+03	2.557e+03	0.309565
Usage	-3.996e+02	1.317e+02	0.002597 **
Prod. attrib	1.107e+01	2.194e+02	0.959810
P. intrection	-3.796e+27	3.587e+27	0.290583
Sel. Crit.	-3.904e+03	1.172e+03	0.000961 ***

$Signif.codes : 0 *** 0.001 ** 0.01 * 0.05.0.11s$

Result of Normal regression cont.

Variable	Estimate	Std. Error	$Pr(>\|t\|)$
Cost	-1.637e+04	2.178e+03	4.93e-13 ***
Terminology	-2.225e+03	3.875e+02	2.07e-08 ***
Positive	5.483e+01	2.205e+02	0.803747
Negative	7.713e+02	3.121e+02	0.013959 *
Comparison	3.178e+03	1.303e+03	0.015273 *
Social NW	-4.046e+03	8.172e+02	1.16e-06 ***
Extreme fl.	-6.964e+02	3.573e+02	0.052117 .
Uncertainty	-2.343e+02	9.585e+02	0.806995
Politeness	-1.079e+04	7.862e+03	0.170868
Broad	4.218e+02	1.099e+03	0.701423
Narrow	1.333e+02	8.181e+01	0.104270

$Signif.codes : 0 * * * 0.001 * *0.01 * 0.05.0.11s$

Result of Normal regression cont.

DF	344
R^2	0.4641
Adjusted R^2	0.4329
p-value	$< 2.2e-16$

Poisson Regression

Variable	Estimate	Std. Error	$Pr(>\|t\|)$
(Interc.)	9.874e+00	1.355e-03	¡ 2e-16 ***
Info Src	6.878e-01	2.526e-02	¡ 2e-16 ***
complement	1.112e-01	3.108e-03	¡ 2e-16 ***
substitute	-2.130e-01	7.821e-03	¡ 2e-16 ***
experience	1.020e-01	7.977e-03	¡ 2e-16 ***
Naive	-1.469e-01	1.703e-02	¡ 2e-16 ***
Usage	-2.169e-02	8.725e-04	¡ 2e-16 ***
Prod. attrib	-7.314e-04	1.466e-03	0.61799
P. intrection	-2.150e+23	2.357e+22	¡ 2e-16 ***

$Signif.codes : 0 * * * 0.001 * *0.01 * 0.05.0.11$

Three Fisher scoring iteration

Poisson Regression cont.

Variable	Estimate	Std. Error	$Pr(>\|t\|)$
Sel. Crit.	-2.128e-01	7.879e-03	¡ 2e-16 ***
Cost	-8.865e-01	1.461e-02	¡ 2e-16 ***
Terminology	-1.229e-01	2.591e-03	¡ 2e-16 ***
Positive	3.737e-03	1.467e-03	0.01085 *
Negative	4.214e-02	2.090e-03	¡ 2e-16 ***
Comparison	1.640e-01	8.625e-03	¡ 2e-16 ***
Social NW	-2.192e-01	5.455e-03	¡ 2e-16 ***
Extreme fl.	-3.872e-02	2.382e-03	¡ 2e-16 ***
Uncertainty	-1.029e-02	6.391e-03	0.10726
Politeness	-5.791e-01	5.177e-02	¡ 2e-16 ***
Broad	2.031e-02	7.290e-03	0.00534 **
Narrow	7.439e-03	5.419e-04	¡ 2e-16 ***

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Three Fisher scoring iteration

Bayesian Gaussian Linear Regression

Variable	Mean	SD
(*Interc.*)	1.9e+04	0.32
Info Src	1.2e+04	0.32
complement	2.0e+03	0.32
substitute	-3.9e+03	0.32
experience	1.8e+03	0.32
Naive	-2.6e+03	0.32
Usage	-4.0e+02	0.32
Prod. attrib	1.1e+01	0.31
P. intrection	-3.8e+27	0.00
Sel. Crit.	-3.9e+03	0.32

$$\text{Probability } (s) = 0.95$$

Bayesian Gaussian Linear Regression cont.

Variable	Mean	SD
Cost	-1.6e+04	0.32
Terminology	-2.2e+03	0.31
Positive	5.5e+01	0.32
Negative	7.7e+02	0.31
Comparison	3.2e+03	0.32
Social NW	-4.0e+03	0.32
Extreme fl.	-7.0e+02	0.32
Uncertainty	-2.3e+02	0.32
Politeness	-1.1e+04	0.31
Broad	4.2e+02	0.32
Narrow	1.3e+02	0.32
Sigma	1231559	89247

Probability $(s) = 0.95$

Bayesian Tobit

Variable	Estimate	SD
(Interc.)	1.939e+04	2.041e+02
Info Src	1.214e+04	3.813e+03
complement	2.042e+03	4.669e+02
substitute	-3.9e+03	0.32
experience	1.779e+03	1.207e+03
Naive	-2.601e+03	2.581e+03
Usage	-3.995e+02	1.322e+02
Prod. attrib	1.007e+01	2.213e+02
P. intrection	-3.797e+27	3.609e+27
Sel. Crit.	-3.908e+03	1.179e+03

Not that much different from Bayesian regression model

Probability $(s) = 0.95$

Bayesian Tobit cont.

Variable	Estimate	SD
Cost	-1.637e+04	2.189e+03
Terminology	-2.224e+03	3.891e+02
Positive	5.621e+01	2.213e+02
Negative	7.698e+02	3.120e+02
Comparison	3.174e+03	1.304e+03
Social NW	-4.048e+03	8.190e+02
Extreme fl.	-6.981e+02	3.584e+02
Uncertainty	-2.341e+02	9.574e+02
Politeness	-1.081e+04	7.871e+03
Broad	44.275e+02	1.100e+03
Narrow	1.332e+02	8.120e+01
Sigma	1.249e+06	9.608e+04

Not that much different from Bayesian regression model

Probability $(s) = 0.95$

Bayesian Ordinary Probit

Variable	Estimate	SD
(Interc.)	2.4602	0.0861
Info Src	0.0249	0.2070
complement	0.0334	0.0402
substitute	0.0205	0.1367
experience	0.0058	0.0705
Naive	0.1426	0.2439
Usage	-0.0744	0.0167
Prod. attrib	-0.0725	0.0301
P. intrection	-0.3579	0.4702
Sel. Crit.	-0.0649	0.2166

Bayesian Ordinary Probit cont.

Variable	Estimate	SD
Cost	0.7268	0.3572
Terminology	0.1061	0.0705
Positive	0.2215	0.0326
Negative	-0.0881	0.0255
Comparison	-0.0509	0.0923
Social NW	0.0279	0.0904
Extreme fl.	0.0910	0.0438
Uncertainty	-0.1971	0.0938
Politeness	0.5938	0.3750
Broad	-0.2004	0.1027
Narrow	-0.0038	0.0072

Bayesian Ordinary Probit cont.

cut off	mean	STD
1	0.00	0.000
2	0.33	0.045
3	0.86	0.071
4	1.78	0.079

Frequentist Binary Probit

Variable	Estimate	SD	$Pr(>\|z\|)$
(Interc.)	-0.933384	-16.468	¡ 2e-16 ***
Info Src	0.266610	0.258385	0.3021
complement	-0.031706	0.046578	0.4961
substitute	0.076687	0.153450	0.6173
experience	0.107060	0.081933	0.1913
Naive	0.014478	0.260034	0.9556
Usage	0.091107	0.021076	1.54e-05 ***
Prod. attrib	-0.030762	0.036012	0.3930
P. intrection	-0.594659	0.609818	0.3295
Sel. Crit.	-0.420550	0.298737	0.1592

$Signif.codes : 0 *** 0.001 ** 0.01 * 0.05.0.11$

Frequentist Binary Probit

Variable	Estimate	SD	$Pr(>\|z\|)$
Terminology	-0.116154	0.083958	0.1665
Positive	-0.174933	0.036411	1.55e-06 ***
Negative	0.014374	0.030931	0.6421
Comparison	-0.104262	0.114742	0.3635
Social NW	-0.170967	0.105617	0.1055
Extreme fl.	0.087655	0.049576	0.0770
Uncertainty	0.218757	0.111400	0.0496 *
Politeness	0.244508	0.435486	0.5745
Broad	-0.035015	0.125350	0.7800
Narrow	0.011943	0.008273	0.1488

$Signif.codes : 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1$

Bayesian Binary Probit

Variable	Estimate	SD
(*Interc.*)	-0.9339	0.0559
Info Src	0.3099	0.2801
complement	-0.0330	0.0456
substitute	0.0607	0.1546
experience	0.1075	0.0821
Naive	-0.0088	0.2680
Usage	0.0934	0.0202
Prod. attrib	-0.0315	0.0360
P. intrection	-0.6773	0.6010
Sel. Crit.	-0.4534	0.3166

Somehow similar to frequentist estimate

Bayesian Binary Probit

Variable	Estimate	SD
Cost	-0.8793	0.5247
Terminology	-0.1181	0.0835
Positive	-0.1795	0.0368
Negative	0.0139	0.0289
Comparison	-0.1134	0.1137
Social NW	-0.1761	0.1048
Extreme fl.	0.0891	0.0508
Uncertainty	0.2204	0.1115
Politeness	0.2663	0.4254
Broad	-0.0353	0.1237
Narrow	0.0114	0.0084

Somehow similar to frequentist estimate

Frequentist Binary Logit

Variable	Estimate	SD	$Pr(>\|z\|)$
(Interc.)	-1.54729	0.10058	¡ 2e-16 ***
Info Src	0.42794	0.44572	0.3370
complement	-0.06649	0.08214	0.4182
substitute	0.16701	0.26707	0.5318
experience	0.19751	0.13781	0.1518
Naive	0.05007	0.45324	0.9120
Usage	0.15809	0.03646	1.45e-05 ***
Prod. attrib	-0.06275	0.06278	0.3175
P. intrection	-1.22210	1.15394	0.2896
Sel. Crit.	-0.67191	0.53827	0.2119

$Signif.codes : 0 *** 0.001 **0.01 * 0.05.0.11$

Frequentist Binary Logit

Variable	Estimate	SD	$Pr(>\|z\|)$
Cost	-1.18657	0.84149	0.1585
Terminology	-0.20762	0.15033	0.1673
Positive	-0.30409	0.06622	4.39e-06 ***
Negative	0.02858	0.05400	0.5966
Comparison	-0.19581	0.20370	0.3364
Social NW	-0.32325	0.18751	0.0847
Extreme fl.	0.15999	0.08616	0.0633 .
Uncertainty	0.38933	0.18779	0.0382 *
Politeness	0.45152	0.73675	0.5400
Broad	-0.06837	0.21904	0.7549
Narrow	0.02104	0.01399	0.1326

$Signif.codes : 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1$

Bayesian Binary Logit

Model worked for low dimension x, but since mine had 20 variable, the code did not work.

Conclusion

1. Bayesian is good, but it generates somehow same result as frequentist approach