**Model 1:**

. logit Outcome IndentationRatio Comments NonDicRatio KeywordRatio SeparatorRati

> o OperatorRatio StopWordRatio

Iteration 0: log likelihood = -290.60953

Iteration 1: log likelihood = -101.51862

Iteration 2: log likelihood = -65.101906

Iteration 3: log likelihood = -52.618033

Iteration 4: log likelihood = -49.221024

Iteration 5: log likelihood = -48.497201

Iteration 6: log likelihood = -48.393144

Iteration 7: log likelihood = -48.388869

Iteration 8: log likelihood = -48.388861

Logistic regression Number of obs = 422

LR chi2(7) = 484.44

Prob > chi2 = 0.0000

Log likelihood = -48.388861 Pseudo R2 = 0.8335

------------------------------------------------------------------------------

Outcome | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

Indentatio~o | .0929289 .1097683 0.85 0.397 -.1222131 .3080709

Comments | 2.703158 1.178144 2.29 0.022 .3940387 5.012278

NonDicRatio | 4.469364 1.773237 2.52 0.012 .9938828 7.944845

KeywordRatio | 6.345914 2.087524 3.04 0.002 2.254442 10.43739

SeparatorR~o | 37.55491 8.466186 4.44 0.000 20.96149 54.14833

OperatorRa~o | 3.106178 6.797098 0.46 0.648 -10.21589 16.42824

StopWordRa~o | -6.189918 2.414507 -2.56 0.010 -10.92226 -1.457571

\_cons | -2.858125 1.137156 -2.51 0.012 -5.08691 -.6293402

------------------------------------------------------------------------------

Note: 0 failures and 6 successes completely determined.

. glm Outcome IndentationRatio Comments NonDicRatio KeywordRatio SeparatorRatio

> OperatorRatio StopWordRatio, family(binomial) link(logit)

Iteration 0: log likelihood = -87.004912

Iteration 1: log likelihood = -51.236757

Iteration 2: log likelihood = -48.841192

Iteration 3: log likelihood = -48.39975

Iteration 4: log likelihood = -48.3889

Iteration 5: log likelihood = -48.388861

Iteration 6: log likelihood = -48.388861

Generalized linear models No. of obs = 422

Optimization : ML Residual df = 414

Scale parameter = 1

Deviance = 96.77772282 (1/df) Deviance = .2337626

Pearson = 212.4571095 (1/df) Pearson = .5131814

Variance function: V(u) = u\*(1-u) [Bernoulli]

Link function : g(u) = ln(u/(1-u)) [Logit]

AIC = .2672458

Log likelihood = -48.38886141 BIC = -2405.854

------------------------------------------------------------------------------

| OIM

Outcome | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

Indentatio~o | .0929289 .1097684 0.85 0.397 -.1222132 .3080709

Comments | 2.703159 1.178145 2.29 0.022 .3940369 5.01228

NonDicRatio | 4.469364 1.773238 2.52 0.012 .9938823 7.944846

KeywordRatio | 6.345914 2.087524 3.04 0.002 2.254442 10.43739

SeparatorR~o | 37.55491 8.466187 4.44 0.000 20.96149 54.14833

OperatorRa~o | 3.106178 6.797098 0.46 0.648 -10.21589 16.42825

StopWordRa~o | -6.189918 2.414507 -2.56 0.010 -10.92227 -1.45757

\_cons | -2.858125 1.137156 -2.51 0.012 -5.08691 -.6293398

**Model 2: Model 1 + Adding word count**

. logit Outcome IndentationRatio Comments NonDicRatio KeywordRatio SeparatorRati

> o OperatorRatio StopWordRatio NumberWords

Iteration 0: log likelihood = -290.60953

Iteration 1: log likelihood = -101.19174

Iteration 2: log likelihood = -64.788106

Iteration 3: log likelihood = -52.227496

Iteration 4: log likelihood = -48.676681

Iteration 5: log likelihood = -47.762925

Iteration 6: log likelihood = -47.507707

Iteration 7: log likelihood = -47.440116

Iteration 8: log likelihood = -47.433421

Iteration 9: log likelihood = -47.433364

Logistic regression Number of obs = 422

LR chi2(8) = 486.35

Prob > chi2 = 0.0000

Log likelihood = -47.433364 Pseudo R2 = 0.8368

------------------------------------------------------------------------------

Outcome | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

Indentatio~o | .1345981 .1045549 1.29 0.198 -.0703258 .339522

Comments | 3.913222 2.514796 1.56 0.120 -1.015687 8.842132

NonDicRatio | 4.194048 1.75011 2.40 0.017 .7638959 7.624201

KeywordRatio | 6.207098 2.095062 2.96 0.003 2.100852 10.31334

SeparatorR~o | 36.05992 8.341631 4.32 0.000 19.71063 52.40922

OperatorRa~o | 2.637451 6.721702 0.39 0.695 -10.53684 15.81174

StopWordRa~o | -5.890858 2.428643 -2.43 0.015 -10.65091 -1.130804

NumberWords | -.0049425 .0049658 -1.00 0.320 -.0146753 .0047904

\_cons | -2.62889 1.120675 -2.35 0.019 -4.825373 -.4324078

------------------------------------------------------------------------------

Note: 1 failure and 13 successes completely determined.

.

. glm Outcome IndentationRatio Comments NonDicRatio KeywordRatio SeparatorRatio

> OperatorRatio StopWordRatio NumberWords, family(binomial) link(logit)

Iteration 0: log likelihood = -86.684734

Iteration 1: log likelihood = -50.801653

Iteration 2: log likelihood = -48.190253

Iteration 3: log likelihood = -47.536918

Iteration 4: log likelihood = -47.438056

Iteration 5: log likelihood = -47.433379

Iteration 6: log likelihood = -47.433364

Iteration 7: log likelihood = -47.433364

Generalized linear models No. of obs = 422

Optimization : ML Residual df = 413

Scale parameter = 1

Deviance = 94.86672807 (1/df) Deviance = .2297015

Pearson = 204.1215667 (1/df) Pearson = .4942411

Variance function: V(u) = u\*(1-u) [Bernoulli]

Link function : g(u) = ln(u/(1-u)) [Logit]

AIC = .2674567

Log likelihood = -47.43336403 BIC = -2401.72

------------------------------------------------------------------------------

| OIM

Outcome | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

Indentatio~o | .1345981 .104555 1.29 0.198 -.070326 .3395222

Comments | 3.913227 2.51491 1.56 0.120 -1.015905 8.842359

NonDicRatio | 4.194048 1.750113 2.40 0.017 .7638908 7.624206

KeywordRatio | 6.207098 2.095069 2.96 0.003 2.100838 10.31336

SeparatorR~o | 36.05992 8.341651 4.32 0.000 19.71059 52.40926

OperatorRa~o | 2.637451 6.721707 0.39 0.695 -10.53685 15.81175

StopWordRa~o | -5.890858 2.428651 -2.43 0.015 -10.65093 -1.13079

NumberWords | -.0049425 .0049659 -1.00 0.320 -.0146755 .0047906

\_cons | -2.62889 1.120677 -2.35 0.019 -4.825376 -.4324043

------------------------------------------------------------------------------

**Model 3: Model 1 + Replacing Comments by LogComments. Not good.**

. glm Outcome IndentationRatio LogComments NonDicRatio KeywordRatio SeparatorRat

> io OperatorRatio StopWordRatio, family(binomial) link(logit)

Iteration 0: log likelihood = -7.5439267

Iteration 1: log likelihood = -6.300e-07

Iteration 2: log likelihood = -6.300e-07

Generalized linear models No. of obs = 63

Optimization : ML Residual df = 55

Scale parameter = 1

Deviance = 1.26000e-06 (1/df) Deviance = 2.29e-08

Pearson = 6.30000e-07 (1/df) Pearson = 1.15e-08

Variance function: V(u) = u\*(1-u) [Bernoulli]

Link function : g(u) = ln(u/(1-u)) [Logit]

AIC = .2539683

Log likelihood = -6.30000e-07 BIC = -227.8724

------------------------------------------------------------------------------

| OIM

Outcome | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

Indentatio~o | -2.950496 998.7549 -0.00 0.998 -1960.474 1954.573

LogComments | 28.29534 1989.27 0.01 0.989 -3870.601 3927.192

NonDicRatio | 40.37547 12156.51 0.00 0.997 -23785.95 23866.7

KeywordRatio | 259.3936 21170.21 0.01 0.990 -41233.45 41752.24

SeparatorR~o | 1165.197 70532.56 0.02 0.987 -137076.1 139406.5

OperatorRa~o | 54.31833 114603.8 0.00 1.000 -224564.9 224673.6

StopWordRa~o | 6.317869 18752.83 0.00 1.000 -36748.56 36761.19

\_cons | -83.4373 10452.14 -0.01 0.994 -20569.26 20402.39

------------------------------------------------------------------------------

**Model 4: Model 2 + Replacing Comments by LogComments. Not good.**

. glm Outcome IndentationRatio LogComments NonDicRatio KeywordRatio SeparatorRat

> io OperatorRatio StopWordRatio NumberWords, family(binomial) link(logit)

Iteration 0: log likelihood = -6.025057

Iteration 1: log likelihood = -6.300e-07

Iteration 2: log likelihood = -6.300e-07

Generalized linear models No. of obs = 63

Optimization : ML Residual df = 54

Scale parameter = 1

Deviance = 1.26000e-06 (1/df) Deviance = 2.33e-08

Pearson = 6.30000e-07 (1/df) Pearson = 1.17e-08

Variance function: V(u) = u\*(1-u) [Bernoulli]

Link function : g(u) = ln(u/(1-u)) [Logit]

AIC = .2857143

Log likelihood = -6.30000e-07 BIC = -223.7293

------------------------------------------------------------------------------

| OIM

Outcome | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

Indentatio~o | 2.638205 1027.831 0.00 0.998 -2011.874 2017.151

LogComments | 6.219285 2013.221 0.00 0.998 -3939.622 3952.06

NonDicRatio | 14.77547 12244.64 0.00 0.999 -23984.29 24013.84

KeywordRatio | 57.23486 22241.78 0.00 0.998 -43535.86 43650.33

SeparatorR~o | 248.4182 72009.09 0.00 0.997 -140886.8 141383.6

OperatorRa~o | 68.89588 115717.9 0.00 1.000 -226733.9 226871.7

StopWordRa~o | -28.19323 18983.14 -0.00 0.999 -37234.47 37178.08

NumberWords | -.0667201 10.18614 -0.01 0.995 -20.03119 19.89775

\_cons | 1.171512 10656.48 0.00 1.000 -20885.14 20887.48

**Model 5: Calculating Probability of being Text**

gen OutcomeText=Outcome

. replace OutcomeText=2 if OutcomeText==1

(191 real changes made)

. replace OutcomeText=1 if OutcomeText==0

(231 real changes made)

. replace OutcomeText=0 if OutcomeText==2

(191 real changes made)

. - preserve

logit OutcomeText KeywordRatio StopWordRatio NumberWords

Iteration 0: log likelihood = -290.60953

Iteration 1: log likelihood = -116.05145

Iteration 2: log likelihood = -88.991742

Iteration 3: log likelihood = -80.458005

Iteration 4: log likelihood = -78.834392

Iteration 5: log likelihood = -78.740078

Iteration 6: log likelihood = -78.739592

Iteration 7: log likelihood = -78.739592

Logistic regression Number of obs = 422

LR chi2(3) = 423.74

Prob > chi2 = 0.0000

Log likelihood = -78.739592 Pseudo R2 = 0.7291

------------------------------------------------------------------------------

OutcomeText | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

KeywordRatio | -12.91307 2.532948 -5.10 0.000 -17.87755 -7.948578

StopWordRa~o | 12.02582 1.29822 9.26 0.000 9.481353 14.57028

NumberWords | .0008251 .0022349 0.37 0.712 -.0035552 .0052053

\_cons | -1.088238 .3410933 -3.19 0.001 -1.756768 -.419707

------------------------------------------------------------------------------

.

.

.

. logit OutcomeText KeywordRatio StopWordRatio

Iteration 0: log likelihood = -290.60953

Iteration 1: log likelihood = -116.25014

Iteration 2: log likelihood = -89.096406

Iteration 3: log likelihood = -80.545266

Iteration 4: log likelihood = -78.921108

Iteration 5: log likelihood = -78.827955

Iteration 6: log likelihood = -78.827505

Logistic regression Number of obs = 422

LR chi2(2) = 423.56

Prob > chi2 = 0.0000

Log likelihood = -78.827505 Pseudo R2 = 0.7288

------------------------------------------------------------------------------

OutcomeText | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

KeywordRatio | -12.97152 2.537638 -5.11 0.000 -17.9452 -7.997839

StopWordRa~o | 12.18541 1.247124 9.77 0.000 9.741092 14.62973

\_cons | -1.06805 .3363529 -3.18 0.001 -1.727289 -.4088102

------------------------------------------------------------------------------

.

.

. logit OutcomeText StopWordRatio

Iteration 0: log likelihood = -290.60953

Iteration 1: log likelihood = -121.16201

Iteration 2: log likelihood = -100.27461

Iteration 3: log likelihood = -97.124628

Iteration 4: log likelihood = -96.945773

Iteration 5: log likelihood = -96.944745

Logistic regression Number of obs = 422

LR chi2(1) = 387.33

Prob > chi2 = 0.0000

Log likelihood = -96.944745 Pseudo R2 = 0.6664

------------------------------------------------------------------------------

OutcomeText | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

StopWordRa~o | 13.12949 1.098094 11.96 0.000 10.97726 15.28171

\_cons | -2.83285 .2838678 -9.98 0.000 -3.389221 -2.27648

------------------------------------------------------------------------------

.

.

. logit OutcomeText StopWordRatio NumberWords

Iteration 0: log likelihood = -290.60953

Iteration 1: log likelihood = -120.8628

Iteration 2: log likelihood = -100.03448

Iteration 3: log likelihood = -96.850541

Iteration 4: log likelihood = -96.65496

Iteration 5: log likelihood = -96.653236

Iteration 6: log likelihood = -96.653236

Logistic regression Number of obs = 422

LR chi2(2) = 387.91

Prob > chi2 = 0.0000

Log likelihood = -96.653236 Pseudo R2 = 0.6674

------------------------------------------------------------------------------

OutcomeText | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

StopWordRa~o | 12.90072 1.129332 11.42 0.000 10.68727 15.11417

NumberWords | .0012085 .0018408 0.66 0.511 -.0023994 .0048165

\_cons | -2.870543 .290769 -9.87 0.000 -3.44044 -2.300646

------------------------------------------------------------------------------

**Only Stats**

**Model 1:**

replace Outcome=1 if Classification=="code"

(298 real changes made)

. replace Outcome=0 if Classification=="text"

(0 real changes made)

. replace Outcome=0 if Classification=="text"

(0 real changes made)

. - preserve

logit Outcome IndentationRatio Comments NonDicRatio KeywordRatio SeparatorRatio

> OperatorRatio StopWordRatio

Iteration 0: log likelihood = -1136.1315

Iteration 1: log likelihood = -1068.9702

Iteration 2: log likelihood = -1035.9608

Iteration 3: log likelihood = -799.25502

Iteration 4: log likelihood = -762.64084

Iteration 5: log likelihood = -755.44616

Iteration 6: log likelihood = -755.41171

Iteration 7: log likelihood = -755.41171

Logistic regression Number of obs = 5112

LR chi2(7) = 761.44

Prob > chi2 = 0.0000

Log likelihood = -755.41171 Pseudo R2 = 0.3351

------------------------------------------------------------------------------

Outcome | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

Indentatio~o | .0032883 .0218021 0.15 0.880 -.039443 .0460196

Comments | 5.776913 .7819965 7.39 0.000 4.244228 7.309598

NonDicRatio | 1.486118 .2063238 7.20 0.000 1.08173 1.890505

KeywordRatio | 2.645179 .21924 12.07 0.000 2.215477 3.074882

SeparatorR~o | 6.002396 .5410827 11.09 0.000 4.941894 7.062899

OperatorRa~o | 1.957991 .7818767 2.50 0.012 .4255408 3.490441

StopWordRa~o | -3.615758 .7407527 -4.88 0.000 -5.067607 -2.16391

\_cons | -3.522055 .1136058 -31.00 0.000 -3.744719 -3.299392

------------------------------------------------------------------------------

Note: 0 failures and 12 successes completely determined.

**Model 2: Model 1 + Adding word count**

. logit Outcome IndentationRatio Comments NonDicRatio KeywordRatio SeparatorRatio OperatorRatio StopWordRatio NumberWords

Iteration 0: log likelihood = -1136.1315

Iteration 1: log likelihood = -1045.0125

Iteration 2: log likelihood = -1001.6323

Iteration 3: log likelihood = -751.37022

Iteration 4: log likelihood = -720.89575

Iteration 5: log likelihood = -717.48624

Iteration 6: log likelihood = -717.23933

Iteration 7: log likelihood = -717.23589

Iteration 8: log likelihood = -717.23589

Logistic regression Number of obs = 5112

LR chi2(8) = 837.79

Prob > chi2 = 0.0000

Log likelihood = -717.23589 Pseudo R2 = 0.3687

------------------------------------------------------------------------------

Outcome | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

Indentatio~o | -.004445 .0240918 -0.18 0.854 -.051664 .042774

Comments | 6.561948 1.010928 6.49 0.000 4.580566 8.543331

NonDicRatio | 1.399523 .2119964 6.60 0.000 .9840177 1.815028

KeywordRatio | 2.532519 .2245135 11.28 0.000 2.09248 2.972557

SeparatorR~o | 5.761614 .5348814 10.77 0.000 4.713266 6.809962

OperatorRa~o | 1.830416 .8247918 2.22 0.026 .2138536 3.446978

StopWordRa~o | -8.287375 1.368831 -6.05 0.000 -10.97023 -5.604515

NumberWords | .0467208 .0063089 7.41 0.000 .0343556 .0590859

\_cons | -3.584177 .1168726 -30.67 0.000 -3.813243 -3.355111

------------------------------------------------------------------------------

Note: 0 failures and 16 successes completely determined.

**Model 5: Calculating Probability of being Text**

. gen OutcomeText=Outcome

. replace OutcomeText=2 if OutcomeText==1

(298 real changes made)

. replace OutcomeText=1 if OutcomeText==0

(4814 real changes made)

. replace OutcomeText=0 if OutcomeText==2

(298 real changes made)

. logit OutcomeText StopWordRatio NumberWords

Iteration 0: log likelihood = -1136.1315

Iteration 1: log likelihood = -1072.063

Iteration 2: log likelihood = -969.56213

Iteration 3: log likelihood = -954.24773

Iteration 4: log likelihood = -950.78816

Iteration 5: log likelihood = -950.52787

Iteration 6: log likelihood = -950.52587

Logistic regression Number of obs = 5112

LR chi2(2) = 371.21

Prob > chi2 = 0.0000

Log likelihood = -950.52587 Pseudo R2 = 0.1634

------------------------------------------------------------------------------

OutcomeText | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

StopWordRa~o | 9.710496 1.09226 8.89 0.000 7.569704 11.85129

NumberWords | -.0623996 .0055703 -11.20 0.000 -.0733172 -.0514821

\_cons | 2.731463 .0677514 40.32 0.000 2.598673 2.864254

http://faculty.chass.ncsu.edu/garson/PA765/logistic.htm

*Logit coefficients, and why they are preferred over odds ratios in modeling*. Note that for the case of decrease the odds ratio can vary only from 0 to .999, while for the case of increase it can vary from 1.001 to infinity. This asymmetry is a drawback to using the odds ratio as a measure of strength of relationship. Odds ratios are preferred for interpretation, but logit coefficients are preferred in the actual mathematics of logistic models. *Warning:* The odds ratio is a different way of presenting the same information as the unstandardized logit (effect) coefficient discussed in the section on logistic regression, and like it, is not recommended when comparing the relative strengths of the independents. The standardized logit (effect) coefficient is used for this purpose.

**Standarizing**

. listcoef, std help

logit (N=5112): Unstandardized and Standardized Estimates

Observed SD: .23432186

Latent SD: 2.8640527

Odds of: 1 vs 0

-------------------------------------------------------------------------------

OutcomeText | b z P>|z| bStdX bStdY bStdXY SDofX

-------------+-----------------------------------------------------------------

StopWordRa~o | 9.71050 8.890 0.000 2.3961 3.3905 0.8366 0.2468

NumberWords | -0.06240 -11.202 0.000 -1.2086 -0.0218 -0.4220 19.3682

-------------------------------------------------------------------------------

b = raw coefficient

z = z-score for test of b=0

P>|z| = p-value for z-test

bStdX = x-standardized coefficient

bStdY = y-standardized coefficient

bStdXY = fully standardized coefficient

SDofX = standard deviation of X

Only PRE

. logit Outcome IndentationRatio Comments NonDicRatio KeywordRatio SeparatorRatio

> OperatorRatio StopWordRatio NumberWords

Iteration 0: log likelihood = -290.60953

Iteration 1: log likelihood = -101.19174

Iteration 2: log likelihood = -64.788106

Iteration 3: log likelihood = -52.227496

Iteration 4: log likelihood = -48.676681

Iteration 5: log likelihood = -47.762925

Iteration 6: log likelihood = -47.507707

Iteration 7: log likelihood = -47.440116

Iteration 8: log likelihood = -47.433421

Iteration 9: log likelihood = -47.433364

Logistic regression Number of obs = 422

LR chi2(8) = 486.35

Prob > chi2 = 0.0000

Log likelihood = -47.433364 Pseudo R2 = 0.8368

------------------------------------------------------------------------------

Outcome | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

Indentatio~o | .1345981 .1045549 1.29 0.198 -.0703258 .339522

Comments | 3.913222 2.514796 1.56 0.120 -1.015687 8.842132

NonDicRatio | 4.194048 1.75011 2.40 0.017 .7638959 7.624201

KeywordRatio | 6.207098 2.095062 2.96 0.003 2.100852 10.31334

SeparatorR~o | 36.05992 8.341631 4.32 0.000 19.71063 52.40922

OperatorRa~o | 2.637451 6.721702 0.39 0.695 -10.53684 15.81174

StopWordRa~o | -5.890858 2.428643 -2.43 0.015 -10.65091 -1.130804

NumberWords | -.0049425 .0049658 -1.00 0.320 -.0146753 .0047904

\_cons | -2.62889 1.120675 -2.35 0.019 -4.825373 -.4324078

------------------------------------------------------------------------------

Note: 1 failure and 13 successes completely determined.

.

.

. listcoef, std help

logit (N=422): Unstandardized and Standardized Estimates

Observed SD: .4983396

Latent SD: 7.5632875

Odds of: 1 vs 0

-------------------------------------------------------------------------------

Outcome | b z P>|z| bStdX bStdY bStdXY SDofX

-------------+-----------------------------------------------------------------

Indentatio~o | 0.13460 1.287 0.198 0.3729 0.0178 0.0493 2.7702

Comments | 3.91322 1.556 0.120 4.7769 0.5174 0.6316 1.2207

NonDicRatio | 4.19405 2.396 0.017 0.7603 0.5545 0.1005 0.1813

KeywordRatio | 6.20710 2.963 0.003 1.0238 0.8207 0.1354 0.1649

SeparatorR~o | 36.05992 4.323 0.000 2.3646 4.7678 0.3126 0.0656

OperatorRa~o | 2.63745 0.392 0.695 0.0708 0.3487 0.0094 0.0268

StopWordRa~o | -5.89086 -2.426 0.015 -1.4014 -0.7789 -0.1853 0.2379

NumberWords | -0.00494 -0.995 0.320 -0.9800 -0.0007 -0.1296 198.2785

-------------------------------------------------------------------------------

b = raw coefficient

z = z-score for test of b=0

P>|z| = p-value for z-test

bStdX = x-standardized coefficient

bStdY = y-standardized coefficient

bStdXY = fully standardized coefficient

SDofX = standard deviation of X

. logit OutcomeText StopWordRatio NumberWords

Iteration 0: log likelihood = -290.60953

Iteration 1: log likelihood = -120.8628

Iteration 2: log likelihood = -100.03448

Iteration 3: log likelihood = -96.850541

Iteration 4: log likelihood = -96.65496

Iteration 5: log likelihood = -96.653236

Iteration 6: log likelihood = -96.653236

Logistic regression Number of obs = 422

LR chi2(2) = 387.91

Prob > chi2 = 0.0000

Log likelihood = -96.653236 Pseudo R2 = 0.6674

------------------------------------------------------------------------------

OutcomeText | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

StopWordRa~o | 12.90072 1.129332 11.42 0.000 10.68727 15.11417

NumberWords | .0012085 .0018408 0.66 0.511 -.0023994 .0048165

\_cons | -2.870543 .290769 -9.87 0.000 -3.44044 -2.300646

------------------------------------------------------------------------------

. listcoef, std help

logit (N=422): Unstandardized and Standardized Estimates

Observed SD: .4983396

Latent SD: 3.6219277

Odds of: 1 vs 0

-------------------------------------------------------------------------------

OutcomeText | b z P>|z| bStdX bStdY bStdXY SDofX

-------------+-----------------------------------------------------------------

StopWordRa~o | 12.90072 11.423 0.000 3.0689 3.5618 0.8473 0.2379

NumberWords | 0.00121 0.657 0.511 0.2396 0.0003 0.0662 198.2785

-------------------------------------------------------------------------------

b = raw coefficient

z = z-score for test of b=0

P>|z| = p-value for z-test

bStdX = x-standardized coefficient

bStdY = y-standardized coefficient

bStdXY = fully standardized coefficient

SDofX = standard deviation of X

<http://www.ats.ucla.edu/stat/Stata/seminars/stata_logistic/default.htm>

Running from New data generated for Machine Learning -- Pre

. clear

. infile IndentationRatio Comments NonDicRatio KeywordRatio SeparatorRatio Operator

> Ratio NumberWords StopWordRatio str4 Classification using "H:\CodeSnippets\PreTra

> inForStata.txt"

(422 observations read)

. gen Outcome=1

. replace Outcome=0 if Classification=="text"

(231 real changes made)

. logit Outcome IndentationRatio Comments NonDicRatio KeywordRatio SeparatorRatio

> OperatorRatio StopWordRatio NumberWords

Iteration 0: log likelihood = -290.60953

Iteration 1: log likelihood = -100.28279

Iteration 2: log likelihood = -63.199531

Iteration 3: log likelihood = -50.310291

Iteration 4: log likelihood = -46.7299

Iteration 5: log likelihood = -45.874089

Iteration 6: log likelihood = -45.642509

Iteration 7: log likelihood = -45.58334

Iteration 8: log likelihood = -45.577835

Iteration 9: log likelihood = -45.57779

Logistic regression Number of obs = 422

LR chi2(8) = 490.06

Prob > chi2 = 0.0000

Log likelihood = -45.57779 Pseudo R2 = 0.8432

------------------------------------------------------------------------------

Outcome | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

Indentatio~o | .1151345 .1149062 1.00 0.316 -.1100776 .3403465

Comments | 3.684229 2.448274 1.50 0.132 -1.1143 8.482758

NonDicRatio | 3.855762 1.720795 2.24 0.025 .4830669 7.228458

KeywordRatio | 5.858271 2.038608 2.87 0.004 1.862672 9.853869

SeparatorR~o | 39.62681 8.821948 4.49 0.000 22.33611 56.91751

OperatorRa~o | 6.200926 7.015306 0.88 0.377 -7.54882 19.95067

StopWordRa~o | -5.832411 2.457419 -2.37 0.018 -10.64886 -1.015957

NumberWords | -.00413 .0050236 -0.82 0.411 -.013976 .005716

\_cons | -2.636576 1.156492 -2.28 0.023 -4.903259 -.3698931

------------------------------------------------------------------------------

Note: 0 failures and 12 successes completely determined.

. gen OutcomeText=Outcome

. replace OutcomeText=2 if OutcomeText==1

(191 real changes made)

. replace OutcomeText=1 if OutcomeText==0

(231 real changes made)

. replace OutcomeText=0 if OutcomeText==2

(191 real changes made)

. - preserve

logit OutcomeText StopWordRatio NumberWords

Iteration 0: log likelihood = -290.60953

Iteration 1: log likelihood = -120.73354

Iteration 2: log likelihood = -99.866913

Iteration 3: log likelihood = -96.697543

Iteration 4: log likelihood = -96.508297

Iteration 5: log likelihood = -96.506747

Iteration 6: log likelihood = -96.506747

Logistic regression Number of obs = 422

LR chi2(2) = 388.21

Prob > chi2 = 0.0000

Log likelihood = -96.506747 Pseudo R2 = 0.6679

------------------------------------------------------------------------------

OutcomeText | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

StopWordRa~o | 12.9596 1.136557 11.40 0.000 10.73199 15.18721

NumberWords | .0009939 .0018346 0.54 0.588 -.0026019 .0045897

\_cons | -2.885624 .2916473 -9.89 0.000 -3.457243 -2.314006

------------------------------------------------------------------------------

Running from New data generated for Machine Learning – Stats Pre

. clear

. infile IndentationRatio Comments NonDicRatio KeywordRatio SeparatorRatio Operator

> Ratio NumberWords StopWordRatio str4 Classification using "H:\CodeSnippets\LineTr

> ainForStata.txt"

(5112 observations read)

. gen Outcome=1

. replace Outcome=0 if Classification=="text"

(4816 real changes made)

. - preserve

logit Outcome IndentationRatio Comments NonDicRatio KeywordRatio SeparatorRatio

> OperatorRatio StopWordRatio NumberWords

Iteration 0: log likelihood = -1130.56

Iteration 1: log likelihood = -1034.104

Iteration 2: log likelihood = -787.23096

Iteration 3: log likelihood = -751.63708

Iteration 4: log likelihood = -749.48

Iteration 5: log likelihood = -695.40882

Iteration 6: log likelihood = -692.72678

Iteration 7: log likelihood = -692.70611

Iteration 8: log likelihood = -692.7061

Logistic regression Number of obs = 5112

LR chi2(8) = 875.71

Prob > chi2 = 0.0000

Log likelihood = -692.7061 Pseudo R2 = 0.3873

------------------------------------------------------------------------------

Outcome | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

Indentatio~o | .0028562 .0239127 0.12 0.905 -.0440118 .0497243

Comments | 6.529787 1.024215 6.38 0.000 4.522363 8.537211

NonDicRatio | 1.904351 .2170783 8.77 0.000 1.478885 2.329816

KeywordRatio | 2.622029 .2276578 11.52 0.000 2.175828 3.06823

SeparatorR~o | 6.248385 .5536096 11.29 0.000 5.163331 7.33344

OperatorRa~o | 2.748342 .6505725 4.22 0.000 1.473243 4.023441

StopWordRa~o | -8.232071 1.404823 -5.86 0.000 -10.98547 -5.478669

NumberWords | .0477158 .0065619 7.27 0.000 .0348547 .060577

\_cons | -3.701997 .1223576 -30.26 0.000 -3.941814 -3.462181

------------------------------------------------------------------------------

Note: 0 failures and 17 successes completely determined.

. gen OutcomeText=Outcome

. replace OutcomeText=2 if OutcomeText==1

(296 real changes made)

. replace OutcomeText=1 if OutcomeText==0

(4816 real changes made)

. replace OutcomeText=0 if OutcomeText==2

(296 real changes made)

. logit OutcomeText StopWordRatio NumberWords

Iteration 0: log likelihood = -1130.56

Iteration 1: log likelihood = -1069.2698

Iteration 2: log likelihood = -965.42844

Iteration 3: log likelihood = -949.78046

Iteration 4: log likelihood = -946.10283

Iteration 5: log likelihood = -945.80134

Iteration 6: log likelihood = -945.79858

Logistic regression Number of obs = 5112

LR chi2(2) = 369.52

Prob > chi2 = 0.0000

Log likelihood = -945.79858 Pseudo R2 = 0.1634

------------------------------------------------------------------------------

OutcomeText | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

StopWordRa~o | 9.86164 1.122897 8.78 0.000 7.660803 12.06248

NumberWords | -.0637526 .005769 -11.05 0.000 -.0750596 -.0524456

\_cons | 2.735239 .0678793 40.30 0.000 2.602198 2.86828

------------------------------------------------------------------------------

.

Sept 20 --- Run with updated data

PRE

. infile IndentationRatio Comments NonDicRatio KeywordRatio SeparatorRatio Operator

> Ratio NumberWords StopWordRatio str4 Classification using "H:\CodeSnippets\Aug20\

> PreTrainForStata.txt"

(422 observations read)

. gen Outcome=1

. replace Outcome=0 if Classification=="text"

(231 real changes made)

. logit Outcome IndentationRatio Comments NonDicRatio KeywordRatio SeparatorRati

> o OperatorRatio StopWordRatio NumberWords

Iteration 0: log likelihood = -290.60953

Iteration 1: log likelihood = -102.3923

Iteration 2: log likelihood = -65.685704

Iteration 3: log likelihood = -53.296394

Iteration 4: log likelihood = -50.073735

Iteration 5: log likelihood = -49.344311

Iteration 6: log likelihood = -49.138904

Iteration 7: log likelihood = -49.091304

Iteration 8: log likelihood = -49.088161

Iteration 9: log likelihood = -49.088149

Logistic regression Number of obs = 422

LR chi2(8) = 483.04

Prob > chi2 = 0.0000

Log likelihood = -49.088149 Pseudo R2 = 0.8311

------------------------------------------------------------------------------

Outcome | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

Indentatio~o | .0855729 .1129572 0.76 0.449 -.1358192 .306965

Comments | 3.443264 2.22831 1.55 0.122 -.9241445 7.810672

NonDicRatio | 3.953067 1.714287 2.31 0.021 .5931259 7.313008

KeywordRatio | 5.872845 1.998722 2.94 0.003 1.955422 9.790268

SeparatorR~o | 40.40647 8.776507 4.60 0.000 23.20484 57.60811

OperatorRa~o | 4.217736 7.063249 0.60 0.550 -9.625978 18.06145

StopWordRa~o | -4.654346 2.367124 -1.97 0.049 -9.293824 -.0148675

NumberWords | -.0043372 .0048594 -0.89 0.372 -.0138615 .0051871

\_cons | -2.716943 1.160054 -2.34 0.019 -4.990607 -.4432789

------------------------------------------------------------------------------

Note: 0 failures and 11 successes completely determined.

.

.

. gen OutcomeText=Outcome

. replace OutcomeText=2 if OutcomeText==1

(191 real changes made)

. replace OutcomeText=1 if OutcomeText==0

(231 real changes made)

. replace OutcomeText=0 if OutcomeText==2

(191 real changes made)

. logit OutcomeText StopWordRatio NumberWords

Iteration 0: log likelihood = -290.60953

Iteration 1: log likelihood = -124.32655

Iteration 2: log likelihood = -104.71253

Iteration 3: log likelihood = -101.97362

Iteration 4: log likelihood = -101.83378

Iteration 5: log likelihood = -101.83283

Iteration 6: log likelihood = -101.83283

Logistic regression Number of obs = 422

LR chi2(2) = 377.55

Prob > chi2 = 0.0000

Log likelihood = -101.83283 Pseudo R2 = 0.6496

------------------------------------------------------------------------------

OutcomeText | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

StopWordRa~o | 12.46287 1.081308 11.53 0.000 10.34354 14.58219

NumberWords | .0011033 .0018204 0.61 0.544 -.0024646 .0046711

\_cons | -2.796973 .2827543 -9.89 0.000 -3.351161 -2.242784

------------------------------------------------------------------------------

.

STATS

.

.

. clear

. infile IndentationRatio Comments NonDicRatio KeywordRatio SeparatorRatio Operat

> orRatio NumberWords StopWordRatio str4 Classification using "H:\CodeSnippets\Au

> g20\LineTrainForStata.txt"

(5112 observations read)

. gen Outcome=1

. replace Outcome=0 if Classification=="text"

(4816 real changes made)

. logit Outcome IndentationRatio Comments NonDicRatio KeywordRatio SeparatorRatio

> OperatorRatio StopWordRatio NumberWords

Iteration 0: log likelihood = -1130.56

Iteration 1: log likelihood = -1034.104

Iteration 2: log likelihood = -787.23096

Iteration 3: log likelihood = -751.63708

Iteration 4: log likelihood = -749.48

Iteration 5: log likelihood = -695.40882

Iteration 6: log likelihood = -692.72678

Iteration 7: log likelihood = -692.70611

Iteration 8: log likelihood = -692.7061

Logistic regression Number of obs = 5112

LR chi2(8) = 875.71

Prob > chi2 = 0.0000

Log likelihood = -692.7061 Pseudo R2 = 0.3873

------------------------------------------------------------------------------

Outcome | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

Indentatio~o | .0028562 .0239127 0.12 0.905 -.0440118 .0497243

Comments | 6.529787 1.024215 6.38 0.000 4.522363 8.537211

NonDicRatio | 1.904351 .2170783 8.77 0.000 1.478885 2.329816

KeywordRatio | 2.622029 .2276578 11.52 0.000 2.175828 3.06823

SeparatorR~o | 6.248385 .5536096 11.29 0.000 5.163331 7.33344

OperatorRa~o | 2.748342 .6505725 4.22 0.000 1.473243 4.023441

StopWordRa~o | -8.232071 1.404823 -5.86 0.000 -10.98547 -5.478669

NumberWords | .0477158 .0065619 7.27 0.000 .0348547 .060577

\_cons | -3.701997 .1223576 -30.26 0.000 -3.941814 -3.462181

------------------------------------------------------------------------------

Note: 0 failures and 17 successes completely determined.

.

. gen OutcomeText=Outcome

. replace OutcomeText=2 if OutcomeText==1

(296 real changes made)

. replace OutcomeText=1 if OutcomeText==0

(4816 real changes made)

. replace OutcomeText=0 if OutcomeText==2

(296 real changes made)

. logit OutcomeText StopWordRatio NumberWords

Iteration 0: log likelihood = -1130.56

Iteration 1: log likelihood = -1069.2698

Iteration 2: log likelihood = -965.42844

Iteration 3: log likelihood = -949.78046

Iteration 4: log likelihood = -946.10283

Iteration 5: log likelihood = -945.80134

Iteration 6: log likelihood = -945.79858

Logistic regression Number of obs = 5112

LR chi2(2) = 369.52

Prob > chi2 = 0.0000

Log likelihood = -945.79858 Pseudo R2 = 0.1634

------------------------------------------------------------------------------

OutcomeText | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

StopWordRa~o | 9.86164 1.122897 8.78 0.000 7.660803 12.06248

NumberWords | -.0637526 .005769 -11.05 0.000 -.0750596 -.0524456

\_cons | 2.735239 .0678793 40.30 0.000 2.602198 2.86828

------------------------------------------------------------------------------

.

.

.

Aug 20 – Correct Data

PRE

. infile IndentationRatio Comments NonDicRatio KeywordRatio SeparatorRatio Operato

> rRatio NumberWords StopWordRatio str4 Classification using "H:\CodeSnippets\Aug2

> 0CorrectData\PreTrainForStata.txt"

(513 observations read)

. gen Outcome=1

. replace Outcome=0 if Classification=="text"

(300 real changes made)

. logit Outcome IndentationRatio Comments NonDicRatio KeywordRatio SeparatorRatio

> OperatorRatio StopWordRatio NumberWords

Iteration 0: log likelihood = -348.17153

Iteration 1: log likelihood = -122.95469

Iteration 2: log likelihood = -66.183778

Iteration 3: log likelihood = -45.012122

Iteration 4: log likelihood = -37.390955

Iteration 5: log likelihood = -35.087325

Iteration 6: log likelihood = -34.69241

Iteration 7: log likelihood = -34.666869

Iteration 8: log likelihood = -34.666186

Iteration 9: log likelihood = -34.666183

Logistic regression Number of obs = 513

LR chi2(8) = 627.01

Prob > chi2 = 0.0000

Log likelihood = -34.666183 Pseudo R2 = 0.9004

------------------------------------------------------------------------------

Outcome | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

Indentatio~o | .1651769 .1871691 0.88 0.378 -.2016679 .5320217

Comments | .3865188 .2964111 1.30 0.192 -.1944362 .9674739

NonDicRatio | -4.477105 1.776811 -2.52 0.012 -7.959591 -.9946187

KeywordRatio | 6.337477 2.230366 2.84 0.004 1.96604 10.70891

SeparatorR~o | 92.89202 15.09983 6.15 0.000 63.2969 122.4871

OperatorRa~o | 10.21649 5.140427 1.99 0.047 .1414389 20.29154

StopWordRa~o | -10.84291 2.929836 -3.70 0.000 -16.58529 -5.100541

NumberWords | -.0047359 .0104212 -0.45 0.650 -.0251611 .0156893

\_cons | -1.509979 .6142772 -2.46 0.014 -2.71394 -.3060179

------------------------------------------------------------------------------

Note: 1 failure and 12 successes completely determined.

.

.

.

. gen OutcomeText=Outcome

. replace OutcomeText=2 if OutcomeText==1

(213 real changes made)

. replace OutcomeText=1 if OutcomeText==0

(300 real changes made)

. replace OutcomeText=0 if OutcomeText==2

(213 real changes made)

. logit OutcomeText StopWordRatio NumberWords

Iteration 0: log likelihood = -348.17153

Iteration 1: log likelihood = -189.94183

Iteration 2: log likelihood = -174.0998

Iteration 3: log likelihood = -172.21704

Iteration 4: log likelihood = -172.1569

Iteration 5: log likelihood = -172.15681

Logistic regression Number of obs = 513

LR chi2(2) = 352.03

Prob > chi2 = 0.0000

Log likelihood = -172.15681 Pseudo R2 = 0.5055

------------------------------------------------------------------------------

OutcomeText | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

StopWordRa~o | 9.971301 .8120954 12.28 0.000 8.379623 11.56298

NumberWords | .0001841 .0010549 0.17 0.861 -.0018834 .0022516

\_cons | -1.705679 .1777066 -9.60 0.000 -2.053977 -1.35738

------------------------------------------------------------------------------

STATS

. clear

. infile IndentationRatio Comments NonDicRatio KeywordRatio SeparatorRatio Operato

> rRatio NumberWords StopWordRatio str4 Classification using "H:\CodeSnippets\Aug2

> 0CorrectData\LineTrainForStata.txt"

(6092 observations read)

. gen Outcome=1

. replace Outcome=0 if Classification=="text"

(5639 real changes made)

. logit Outcome IndentationRatio Comments NonDicRatio KeywordRatio SeparatorRatio

> OperatorRatio StopWordRatio NumberWords

Iteration 0: log likelihood = -1612.9981

Iteration 1: log likelihood = -1434.4639

Iteration 2: log likelihood = -1269.0182

Iteration 3: log likelihood = -1202.7977

Iteration 4: log likelihood = -1191.2994

Iteration 5: log likelihood = -1190.3941

Iteration 6: log likelihood = -1190.3656

Iteration 7: log likelihood = -1190.3655

Logistic regression Number of obs = 6092

LR chi2(8) = 845.27

Prob > chi2 = 0.0000

Log likelihood = -1190.3655 Pseudo R2 = 0.2620

------------------------------------------------------------------------------

Outcome | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

Indentatio~o | -.0397176 .0228292 -1.74 0.082 -.084462 .0050268

Comments | 3.277092 .4939874 6.63 0.000 2.308894 4.245289

NonDicRatio | 1.38653 .1835911 7.55 0.000 1.026698 1.746362

KeywordRatio | 2.318349 .1741378 13.31 0.000 1.977045 2.659653

SeparatorR~o | 3.06378 .2959273 10.35 0.000 2.483773 3.643787

OperatorRa~o | 1.398545 .6155412 2.27 0.023 .1921067 2.604984

StopWordRa~o | -8.308202 1.100731 -7.55 0.000 -10.46559 -6.15081

NumberWords | .0420243 .0054932 7.65 0.000 .0312578 .0527907

\_cons | -2.931686 .0795605 -36.85 0.000 -3.087621 -2.77575

------------------------------------------------------------------------------

Note: 0 failures and 8 successes completely determined.

.

.

.

. gen OutcomeText=Outcome

. replace OutcomeText=2 if OutcomeText==1

(453 real changes made)

. replace OutcomeText=1 if OutcomeText==0

(5639 real changes made)

. replace OutcomeText=0 if OutcomeText==2

(453 real changes made)

. logit OutcomeText StopWordRatio NumberWords

Iteration 0: log likelihood = -1612.9981

Iteration 1: log likelihood = -1496.6308

Iteration 2: log likelihood = -1425.6358

Iteration 3: log likelihood = -1409.0518

Iteration 4: log likelihood = -1405.2901

Iteration 5: log likelihood = -1404.9795

Iteration 6: log likelihood = -1404.9766

Logistic regression Number of obs = 6092

LR chi2(2) = 416.04

Prob > chi2 = 0.0000

Log likelihood = -1404.9766 Pseudo R2 = 0.1290

------------------------------------------------------------------------------

OutcomeText | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

StopWordRa~o | 10.47461 1.09427 9.57 0.000 8.329885 12.61934

NumberWords | -.0601182 .0054212 -11.09 0.000 -.0707436 -.0494929

\_cons | 2.390988 .0532494 44.90 0.000 2.286621 2.495355

------------------------------------------------------------------------------

.

.

.