Logistic Regression Example

Dave Lorenz

July 26, 2017

Abstract

This examples demonstrates the `binaryReg` and other logistic regression support functions in the `smwrStats` package. The example uses the PugetNitrate dataset from Tesoriero and Voss (1997). The dataset is available from the `smwrData` package.

Contents

1 Introduction

These examples use data from the smwrData package. The data are retrieved in the following code.

```
> # Load the smwrStats and smwrData packages
> library(smwrStats)
> library(smwrData)
> # Get the dataset
> data(PugetNitrate)
> head(PugetNitrate)

  wellid       l10       l20       l40  surfgeo       date nitrate
1   1000 15.375154  0.000000 57.687577   Coarse 1990-09-06     0.2
2   1001  7.839196 77.185930  9.849246   Coarse 1993-06-17     9.4
3   1002  7.236181 35.276382 53.969849   Coarse 1991-05-14     0.4
4   1003 34.472362 11.155779 53.668342   Coarse 1992-05-11     1.0
5   1004  4.623116 13.869347 81.507538 Alluvium 1989-03-17     0.2
6   1005 54.974874  0.201005 21.507538   Coarse 1988-09-19     2.8
   wellmet
1  60.9600
2   5.4864
3  21.9456
4 113.9952
5  30.1752
6  16.7640
```

2 Single Variable Model

The `hosmerLemeshow.test`, `leCessie.test`, and `roc` functions performs diagnostic tests on a logistic regression model created by `glm`. The model can be constructed from either discrete values or counts of successes and failures.

This example follows the assumptions in Tesoriero and Voss (1997). The regression will model the probability that the concentration equals or exceeds 3 mg/L as was done in that report. This example demonstrates the `hosmerLemeshow.test` and `roc` functions on one single variable model described by Tesoriero and Voss (1997). The `leCessie.test` is useful for glm models with fewer than 1000 observations because of the time required to process larger sample sizes.

```
> # Create the logistic regression model
> PSNO3.1 <- glm(formula = nitrate >= 3 ~ wellmet, family = binomial,
+     data = PugetNitrate,  na.action = na.exclude)
> # Print the summary
> print(summary(PSNO3.1))

Call:
glm(formula = nitrate >= 3 ~ wellmet, family = binomial, data = PugetNitrate,
    na.action = na.exclude)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7066  -0.4635  -0.3338  -0.1904   3.0984

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.224334   0.161778  -7.568 3.79e-14 ***
wellmet     -0.029482   0.003857  -7.644 2.10e-14 ***
---
Signif. codes:  0 Ś***Š 0.001 Ś**Š 0.01 Ś*Š 0.05 Ś.Š 0.1 Ś Š 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1014.85  on 1966  degrees of freedom
Residual deviance:  925.19  on 1965  degrees of freedom
AIC: 929.19

Number of Fisher Scoring iterations: 7
```

The statistics from the printed summary agree reasonably well with table 2 in Tesoriero and Voss (1997). Small differences can be expected among different logistic regression implementations due to differences in convergence criteria for example. The G statistics in table 2 is the difference between the null deviance and the model deviance, 1014.85 - 925.19 = 89.66.

The `hosmerLemeshow.test` can help diagnose lack of fit and the output can help construct diagnostic plots like figure 2 in Tesoriero and Voss (1997). The code below runs the test and creates a graph to replicate figure 2, very small differences can be noted due to small differences in grouping.

```
> # Run the H-L test
> PSNO3.1.hl <- hosmerLemeshow.test(PSNO3.1)
> print(PSNO3.1.hl)

        Hosmer-Lemeshow goodness of fit test

data:  nitrate >= 3 ~ wellmet
Chi-square = 22.437, Number of groups = 10, p-value = 0.004167
alternative hypothesis: Some lack of fit
null hypothesis: No lack of fit
sample estimates:
   Size Expected Counts   wellmet
1   196    0.751      1 172.67231
2   199    2.965      3 101.52597
3   193    4.917     12  82.38760
4   206    8.104      8  67.11370
5   191   10.476      5  55.11933
6   188   12.848     14  47.14186
7   203   17.736     10  38.15706
8   199   21.979     16  29.28531
9   196   26.677     28  21.22714
10  196   34.547     44  10.89038

> # Added fitted values to dataset for line in figure 2, and order
> PugetNitrate$fits <- fitted(PSNO3.1)
> OrderFits <- order(PugetNitrate$fits)
> # setSweave is a specialized function that sets up the graphics page for
> # Sweave scripts. For interactive use, it should be removed and the
> # default setting for set.up can be used.
> setSweave("binplot01", 5, 5)
> with(PugetNitrate, xyPlot(wellmet[OrderFits], fits[OrderFits],
+     Plot=list(what="lines"),
+     xaxis.range=c(0, 200),
+     yaxis.range=c(0, .25),
+     xtitle="Well Depth, in meters",
+     ytitle="Estimated Pobability"))
> # Add the observed frequencies
```

```
> with(PSNO3.1.hl$estimate, addXY(wellmet, Counts/Size,
+     Plot=list(what="points")))
> # Required call to close PDF output graphics
> graphics.off()
```

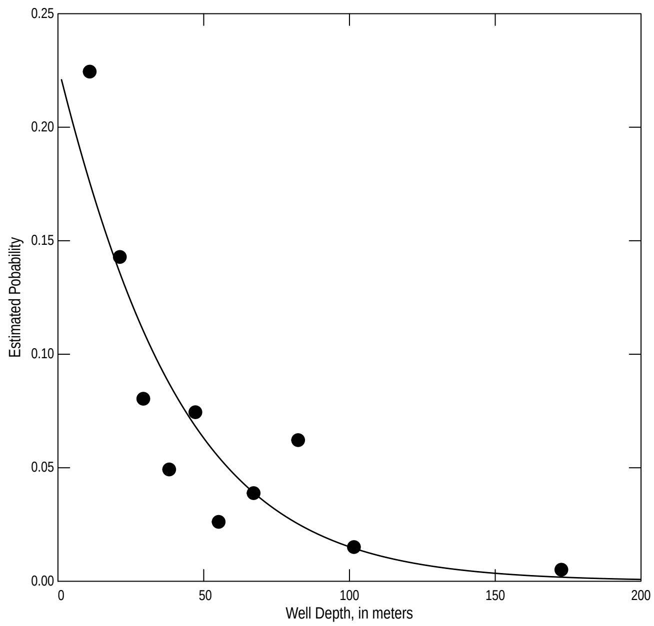

Figure 1. The estimated pobability that nitrate exceeds 3 mg/L as a function of well depth.

The Hosmer-Lemeshow test can be very sensitive to the number of groups. Compare the p-values from the previous test using the default 10 groups with the output below for 12 groups.

```
> # Run the H-L test with 12 groups
> hosmerLemeshow.test(PSNO3.1, 12)

        Hosmer-Lemeshow goodness of fit test

data:  nitrate >= 3 ~ wellmet
Chi-square = 15.603, Number of groups = 12, p-value = 0.1116
alternative hypothesis: Some lack of fit
null hypothesis: No lack of fit
```

```
sample estimates:
   Size Expected Counts    wellmet
1   162    0.466      0 183.632593
2   162    1.942      3 109.071363
3   171    3.567      7  89.258274
4   160    4.906      7  75.763755
5   166    7.160      7  63.725234
6   164    9.171      5  54.388215
7   162   10.901     10  47.688030
8   172   14.207     12  40.208791
9   157   15.984      9  32.365101
10  160   19.137     21  26.216610
11  173   24.963     22  18.911695
12  158   28.596     38   9.761316
```

Another quick evaluation of a logisitc regression is the area under the receiver-operating-curve (AUROC). It is a measure of the predictive power of the model. The result is a number from varies from 0.5, no predicitve power, to 1.0, perfect prediction. Tape, from `http://gim.unmc.edu/dxtests/Default.htm`, accessed on 01/27/2009, gives general guidelines for the AUROC: .50-.60, fail; .60-70, poor; .70-80, fair, .80-.90 good, and .90-1.00 excellent. The `roc` function computes the statistic for any model. The output from the single variable model is shown below. The result indicates fair prediciton.

```
> # Compute the area under the ROC
> roc(PSNO3.1)

Area under the ROC curve: 0.732
```

3 Multiple Variable Model

The binaryReg function performs a series of diagnostic tests on a logistic regression model created by glm. The model can be constructed from either discrete values or counts of successes and failures.

This example follows the assumptions in Tesoriero and Voss (1997) for the groundwater vulnerability model for coarse-grained glacial materials. The regression will model the probability that the concentration equals or exceeds 3 mg/L as was done in that report. This example demonstrates the binaryReg function.

```
> # Create the logistic regression model
> PSNO3.3 <- glm(formula = nitrate >= 3 ~ wellmet + l20 + l10,
+     family = binomial, subset = surfgeo == "Coarse",
+     data = PugetNitrate,  na.action = na.omit)
> # Create the assessment and print it
> PSNO3.3.br <- binaryReg(PSNO3.3)
> print(PSNO3.3.br)
```

```
Call:
glm(formula = nitrate >= 3 ~ wellmet + l20 + l10, family = binomial,
    data = PugetNitrate, subset = surfgeo == "Coarse", na.action = na.omit)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5005  -0.4720  -0.3274  -0.1869   3.0998

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.067279   0.340674  -6.068 1.29e-09 ***
wellmet     -0.028260   0.005854  -4.827 1.38e-06 ***
l20          0.033697   0.006033   5.586 2.33e-08 ***
l10          0.029039   0.006281   4.624 3.77e-06 ***
---
Signif. codes:  0 Ś***Š 0.001 Ś**Š 0.01 Ś*Š 0.05 Ś.Š 0.1 Ś Š 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 518.48  on 718  degrees of freedom
Residual deviance: 409.71  on 715  degrees of freedom
  (23 observations deleted due to missingness)
AIC: 417.71

Number of Fisher Scoring iterations: 6
```

```
Likelihood ratio test: 108.772 on 3 degrees of freedom, p-value is 0

Response profile:
  nitrate >= 3 Response Counts
1        FALSE        0    635
2         TRUE        1     84

 Goodness of fit tests

        le Cessie-van Houwelingen GOF test

data:  nitrate >= 3 ~ wellmet + l20 + l10
Chisq = 22.876, df = 13.509, p-value = 0.0523
alternative hypothesis: Some lack of fit
null hypothesis: No lack of fit
sample estimates:
       Q     E[Q]    se[Q]
58.56150 34.58332 13.30655

Distance between observations:
  maximum bandwidth
 6.237748  1.471405


        Hosmer-Lemeshow goodness of fit test

data:  nitrate >= 3 ~ wellmet + l20 + l10
Chi-square = 1.7, Number of groups = 10, p-value = 1
alternative hypothesis: Some lack of fit
null hypothesis: No lack of fit
sample estimates:
   Size Expected Counts
1    72    0.460      1
2    72    1.408      2
3    72    2.329      2
4    72    3.326      3
5    72    4.335      4
6    71    5.612      5
7    72    7.332      7
8    72    9.566      8
9    72   14.846     15
10   72   34.786     37


Predictive power estimates:
```

```
McFadden R-squared: 0.2098
adjusted R-squared: 0.1982


Classification table.
Percent correct: (1 is sensitivity, 0 is specificity)
   1    0
25.0 97.8

Concordance Index, based on 53340 pairs
Discordant       Tied Concordant
 18.830146   0.001875  81.167979

Area under the ROC curve: 0.812


Influence diagnostic test criteria:
leverage   cooksD    dfits
 0.02086  0.89220  0.34745
        Observations exceeding at least one test criterion
     X...nitrate.X3   yhat  resids deviance.res pearson.res  leverage
2              TRUE 0.6471  0.3529       0.9330      0.7385 0.026464*
16             TRUE 0.3369  0.6631       1.4752      1.4030 0.009688
70            FALSE 0.6556 -0.6556      -1.4600     -1.3796 0.025465*
209           FALSE 0.6308 -0.6308      -1.4117     -1.3071 0.026157*
324            TRUE 0.5081  0.4919       1.1637      0.9839 0.041930*
345            TRUE 0.4948  0.5052       1.1862      1.0104 0.016866
465           FALSE 0.4309 -0.4309      -1.0618     -0.8701 0.024294*
475           FALSE 0.6252 -0.6252      -1.4010     -1.2916 0.038238*
503            TRUE 0.6516  0.3484       0.9256      0.7312 0.025533*
564            TRUE 0.5712  0.4288       1.0584      0.8665 0.021289*
578           FALSE 0.6716 -0.6716      -1.4923     -1.4300 0.027909*
584           FALSE 0.5343 -0.5343      -1.2362     -1.0711 0.022086*
599           FALSE 0.5801 -0.5801      -1.3174     -1.1754 0.022359*
643           FALSE 0.3427 -0.3427      -0.9161     -0.7220 0.021726*
687            TRUE 0.6792  0.3208       0.8795      0.6872 0.030449*
710           FALSE 0.3150 -0.3150      -0.8699     -0.6781 0.009529
732           FALSE 0.6756 -0.6756      -1.5005     -1.4431 0.024312*
733            TRUE 0.6718  0.3282       0.8920      0.6990 0.024399*
734           FALSE 0.6545 -0.6545      -1.4579     -1.3763 0.024823*
1106          FALSE 0.6027 -0.6027      -1.3587     -1.2317 0.021197*
1149           TRUE 0.6069  0.3931       0.9994      0.8048 0.023333*
1298           TRUE 0.5932  0.4068       1.0220      0.8282 0.025341*
1302          FALSE 0.6519 -0.6519      -1.4527     -1.3683 0.024970*
1407          FALSE 0.3451 -0.3451      -0.9202     -0.7260 0.011029
1429           TRUE 0.6115  0.3885       0.9917      0.7970 0.029121*
```

```
1499          FALSE 0.4160 -0.4160       -1.0372     -0.8440 0.032769*
1517           TRUE 0.4799  0.5201        1.2118      1.0411 0.038195*
1518          FALSE 0.4863 -0.4863       -1.1542     -0.9730 0.038610*
1524          FALSE 0.5722 -0.5722       -1.3032     -1.1566 0.024865*
1535           TRUE 0.6894  0.3106        0.8625      0.6713 0.026537*
1628          FALSE 0.5952 -0.5952       -1.3448     -1.2125 0.025310*
1629           TRUE 0.6558  0.3442        0.9187      0.7245 0.031254*
1748          FALSE 0.3710 -0.3710       -0.9630     -0.7680 0.032507*
1775          FALSE 0.1171 -0.1171       -0.4991     -0.3642 0.022628*
1776           TRUE 0.4444  0.5556        1.2736      1.1181 0.040658*
1777          FALSE 0.1137 -0.1137       -0.4913     -0.3582 0.022933*
1780          FALSE 0.1486 -0.1486       -0.5672     -0.4178 0.025516*
1781           TRUE 0.3834  0.6166        1.3847      1.2683 0.037391*
1782          FALSE 0.2802 -0.2802       -0.8109     -0.6240 0.030746*
1850          FALSE 0.4639 -0.4639       -1.1166     -0.9302 0.038909*
1904           TRUE 0.5667  0.4333        1.0658      0.8745 0.022855*
1935           TRUE 0.3890  0.6110        1.3741      1.2532 0.010635
        cooksD      dfits
2    4.819e-02  -0.440932*
16   3.310e-02   0.367118*
70   7.237e-02  -0.541882*
209  2.313e-02  -0.304688
324  6.783e-03  -0.164669
345  3.249e-02   0.362151*
465  1.002e-02   0.200280
475  1.303e-02  -0.228315
503  6.426e-02  -0.510145*
564  1.799e-03   0.084783
578  1.109e-01  -0.672834*
584  2.076e-02   0.288699
599  1.255e-04   0.022392
643  1.585e-02   0.252151
687  1.484e-01  -0.780348*
710  3.004e-02   0.349483*
732  1.232e-01  -0.711441*
733  1.123e-01  -0.678452*
734  6.962e-02  -0.531418*
1106 4.581e-03  -0.135351
1149 4.907e-03  -0.140079
1298 2.246e-04  -0.029952
1302 6.156e-02  -0.499263*
1407 4.081e-02   0.407956*
1429 9.471e-03  -0.194669
1499 7.507e-03   0.173268
1517 1.975e-05  -0.008882
1518 3.318e-04  -0.036405
```

```
1524 3.405e-03   0.116667
1535 1.788e-01  -0.861079*
1628 2.732e-04  -0.033035
1629 7.783e-02  -0.561383*
1748 7.957e-03   0.178396
1775 1.879e-02  -0.274550
1776 1.497e-03   0.077341
1777 1.880e-02  -0.274683
1780 1.562e-02  -0.250167
1781 6.383e-03   0.159745
1782 3.558e-04   0.037698
1850 3.632e-04   0.038091
1904 4.133e-03   0.128556
1935 3.819e-02   0.394507*
```

References

[1] Tesoriero, A.J., and Voss, F.D., 1997, Predicting the probability of elevated nitrate concentrations in the Puget Sound Basin???Implications for aquifer susceptibility and vulnerability: Groundwater, v. 35, no. 6, p. 1029???1039.

[2] Helsel, D.R. and Hirsch, R.M., 2002, Statistical methods in water resources: U.S. Geological Survey Techniques of Water-Resources Investigations, book 4, chap. A3, 522 p.