STAC51 TUT02

Week 12: Apr 01, 2021

Consider the data on two variables: x (sell width of a female crab) and y (the number of satellites)below. Use these codes for R.

x <- c(28.3, 22.5, 26, 24.8, 26, 23.8)

y <- c(8, 0, 9, 0, 4, 0)

Fit a Poisson regression model for this data with a log link function. Use the likelihood ratio test for testing the null hypothesis: $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$. What is the test statistic value? Round the number to the second decimal place, e.g. 3.4786 would be **3.48**.

True or False?

If X and Y are binary, and Z has K categories, so the data can be summarized in a 2 x 2 x K contingency table, one can test conditional independence of X and Y, controlling for Z, using a Wald test or likelihood ratio test of $H_0 : \beta = 0$ in the model

$$logit(P(Y = 1)) = \alpha + \beta x + \beta_1 z_1 + \ldots \beta_{K-1} z_{K-1}$$

where $z_i = 1$ for observations in category *i* of Z and $z_i = 0$ otherwise.

Can I use this model to test homogenous association between X and Y?
Think about conditional independence vs homogenous association.

The table below displays data come from a study on urinary tract infections. Investigators applied two treatments to patients who had wither complicated cases of urinary tract infections are difficult to cure, investigators were interested in whether the pattern of treatment differences are the same across diagnoses.

Diagnosis	Treatment	Cured	Not Cured
Complicated	A	78	28
Complicated	B	101	11
Uncomplicated	A	40	5
Uncomplicated	B	54	6

logit(pi_y) = trt + Diag
logit(pi_y) = trt

Use the R codes below:

```
CuredYes = c(78, 101, 40, 54)
CuredNo = c(28, 11, 5, 6)
TRT = c("A","B", "A", "B")
Diag = c(rep("Complicated",2), rep("Uncomplicated", 2))
data.frame(Diag, TRT, CuredYes, CuredNo)
```

Provide an estimate of the common odds ratio between being cured and treatement controlled by the diagnoses. Round the number to the first decimal place, e.g. 3.4786 would be **3.5**.

The table below displays data come from a study on urinary tract infections. Investigators applied two treatments to patients who had wither complicated cases of urinary tract infections are difficult to cure, investigators were interested in whether the pattern of treatment differences are the same across diagnoses.

Diagnosis	Treatment	Cured	Not Cured
Complicated	A	78	28
Complicated	B	101	11
Uncomplicated	A	40	5
Uncomplicated	B	54	6

Use the R codes below:

```
CuredYes = c(78, 101, 40, 54)
CuredNo = c(28, 11, 5, 6)
TRT = c("A","B", "A", "B")
Diag = c(rep("Complicated",2), rep("Uncomplicated", 2))
data.frame(Diag, TRT, CuredYes, CuredNo)
```

Fit a logistic regression model with the main effect of TRT and Diag, but no interaction.

Test the goodness of fit based on Deviance. Use the level of significance as 0.05. Which one is the correct statement?

- P-value is 0.149, and the homogeneous association model fits the data well.
- P-value is less than 5%, so the homogenous association model fits the data well.
- P-value is 0.851, so the homogeneous association model does have lack-of fit.
- P-value is less than 5%, so the the homogeneous association model does have lack of fit.
- P-value is 0.851, so the homogeneous association model fits the data well.

The table below displays data come from a study on urinary tract infections. Investigators applied two treatments to patients who had wither complicated cases of urinary tract infections are difficult to cure, investigators were interested in whether the pattern of treatment differences are the same across diagnoses.

Diagnosis	Treatment	Cured	Not Cured
Complicated	A	78	28
Complicated	B	101	11
Uncomplicated	A	40	5
Uncomplicated	B	54	6

Use the R codes below:

```
CuredYes = c(78, 101, 40, 54)
CuredNo = c(28, 11, 5, 6)
TRT = c("A","B", "A", "B")
Diag = c(rep("Complicated",2), rep("Uncomplicated", 2))
data.frame(Diag, TRT, CuredYes, CuredNo)
```

Report the upper limit of the 95% **likelihood ratio confidence interval** for the common odds ratio between being cured and treatment controlling for the diagnoses. Round the number to the second decimal place, e.g. 3.4786 would be **3.48**.

True or False?

For a sample of retired subjects in Florida, a contingency table is used to relate X = cholesterol (6 ordered levels) to Y = whether the subject has symptoms of heart disease (yes = 1, no = 0).

For the logistic regression model, $logit[P(Y = 1)] = \alpha + \beta x$ fitted in the 6× 2 contingency table by assigning scores to the 6 cholesterol levels, the deviance statistic equals **3.0**. Thus, this model provides a poor fit to the data. The level of significance is 0.05.

- True
- False

The data come from a study on urinary tract infections. The investigators applied **three treatments** to patients who had either a **complicated or uncomplicated diagnosis** of urinary tract infections. The response variable is **cured or not cured**. Software reports model -2log likelihood values of **492.029** with only an intercept term, **450.071**, with main effect predictors, and **447.556** with all the two-factor interactions. Compute AIC for these three models, and which model among the three models is preferable?

- The main effect model, since it has the smallest AIC of 458.071
- The interaction model, since it has the smallest AIC of 459.556
- The main effect model, since it has the smallest AIC of 459.556
- The interaction model, since it has the smallest AIC of 458.071
- The only intercept model, since it has the largest AIC of 494.029

AIC: 2k - 2log(L)
Null model: 2 * 1 + 492.029 = 494.029
The main effect model: 2* (1 + 2 + 1) + 450.071 = 458.071. logit(pi_y) = intercept + b1 * x1 + b2 * x2 + b3 * Z
The two factor interaction model: 2 * (1 + 3 + 2) + 447.556 = 459.556

For this problem you need to load the NHANES dataset using the following command

```
## If the package is not installed then use ##
install.packages('NHANES') ## And install.packages('tidyverse')
library(tidyverse)
library(NHANES)
small.nhanes <- na.omit(NHANES[NHANES$SurveyYr=="2009_10"
& NHANES$Age > 17,c(1,3,4,7, 9:11,13,25,61)])
small.nhanes <- small.nhanes %>%
group_by(ID) %>% filter(row_number()==1)
```

This is data collected by US National Center for Health Statistics (NCHS). The preceeding codes creates a small dataset of the original NHANES dataset. With this dataset answer the following questions,

(a) Randomly select 500 observations from the data. For this selection use your student ID as seed. Fit a logistic regression to predict smoking status (variable `SmokeNow`), using all the other variables (excluding `ID`). Explain your results in few sentences. [10 Marks]

(b) Perform model selection using AIC/BIC based stepwise approach. [10 Marks]

(c) Perform an internal validation using cross-validation. Explain your results. [10 Marks]

(d) Construct the Receiver operating characteristic (ROC) curve. Calculate the area under the curve (AUC). How would you interpret the AUC. [10 Marks]

(e) Predict the probabilities for the remaining 476 observations. Calculate the deciles for the predicted probabilities. Does the observed and the predicted probabilities differ for the deciles? [10 Marks]

			Primary Food Choice				
Lake	*Sex*	*Size*	*Fish*	*Inv.*	*Rept.*	*Bird*	*Other*
Hancock	M	small	7	1	0	0	5
		large	4	0	0	1	2
	F	small	16	3	2	2	3
		large	3	0	1	2	3
Oklawaha	M	small	2	2	0	0	1
		large	13	7	6	0	0
	F	small	3	9	1	0	2
		large	0	1	0	1	0
Trafford	M	small	3	7	1	0	1
		large	8	6	6	3	5
	F	small	2	4	1	1	4
		large	0	1	0	0	0
George	M	small	13	10	0	2	2
		large	9	0	0	1	2
	F	small	3	9	1	0	1
		large	8	1	0	0	1

the usual primary food choice of alligators appears to be fish, we'll use *fish as the baseline category*

We let

π_1 = prob. of fish,

π_2 = prob. of invertebrates,

π_3 = prob. of reptiles,

π_4 = prob. of birds,

π_5 = prob. of other,

and make "fish" be the baseline category. The logit equations are

$$\log\left(\frac{\pi_j}{\pi_1}\right) = \beta_0 + \beta_1 X_1 + \cdots$$

for $j = 2, 3, 4, 5$. The X's included

- three dummy indicators for lake,
- a dummy for sex, and
- a dummy for size.

Therefore, each logit equation had six coefficients to be estimated, so the number of free parameters in this model is $4 \times 6 = 24$.

```
Call:
vglm(formula = cbind(Bird, Invertebrate, Reptile, Other, Fish) ~
    Lake + Size + Gender, family = multinomial, data = gator)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept):1  -2.4633     0.7739      NA       NA
(Intercept):2  -2.0745     0.6117  -3.392 0.000695 ***
(Intercept):3  -2.9141     0.8856      NA       NA
(Intercept):4  -0.9167     0.4782  -1.917 0.055217 .
Lakegeorge:1   -0.5753     0.7952  -0.723 0.469429
Lakegeorge:2    1.7805     0.6232   2.857 0.004277 **
Lakegeorge:3   -1.1295     1.1928  -0.947 0.343687
Lakegeorge:4   -0.7666     0.5686  -1.348 0.177563
Lakeoklawaha:1 -1.1256     1.1923  -0.944 0.345132
Lakeoklawaha:2  2.6937     0.6693   4.025 5.70e-05 ***
Lakeoklawaha:3  1.4008     0.8105   1.728 0.083926 .
Lakeoklawaha:4 -0.7405     0.7421  -0.998 0.318372
Laketrafford:1  0.6617     0.8461   0.782 0.434145
Laketrafford:2  2.9363     0.6874   4.272 1.94e-05 ***
Laketrafford:3  1.9316     0.8253   2.340 0.019263 *
Laketrafford:4  0.7912     0.5879   1.346 0.178400
Size>2.3:1      0.7302     0.6523   1.120 0.262918
Size>2.3:2     -1.3363     0.4112  -3.250 0.001155 **
Size>2.3:3      0.5570     0.6466   0.861 0.388977
Size>2.3:4     -0.2906     0.4599  -0.632 0.527515
Genderf:1       0.6064     0.6888   0.880 0.378666
Genderf:2       0.4630     0.3955   1.171 0.241796
Genderf:3       0.6276     0.6853   0.916 0.359785
Genderf:4       0.2526     0.4663   0.542 0.588100
---
```

That is the equation for the log odds of food type food={Bird, Inve, Other, Rept} relative to fish (F) for lake (l), size (s) and sex (g):

$$\log \frac{\pi_{food,lsg}}{\pi_{Flsg}} = \alpha_{food} + \beta^{L}_{food,l} + \beta^{S}_{food,s} + \beta^{G}_{food,g}$$

where $\alpha_F = 0, \beta_{food,Hancock} = \beta_{food,small} = \beta_{food,male} = 0$ for $food = Fish, Bird, Inve, Other, Rept$, indicating the reference levels.

For example, the estimated prediction equation for the log-odds of ***birds relative to fish***:

$$log\frac{\hat{\pi}_{Bird,lsg}}{\hat{\pi}_{Fish,lsg}} = -2.4633 + -1.1256Oklawaha + 0.6617Trafford - 0.5753George + 0.7302large + 0.6064female$$

The intercepts give the estimated log-odds for the reference group lake = Hancock, size = small, sex = male. For example, the estimated log-odds of birds versus fish in this group is −2.4633; the estimated log-odds of invertebrates versus fish is −2.0744; and so on.

The lake effect is characterized by three dummy coefficients in each of the four logit equations. The estimated coefficient for the Lake Oklawaha dummy in the bird-versus-fish equation is −1.1256 with st. error 1.1924. This means that alligators in Lake Oklawaha are less likely to choose birds over fish than their colleagues in Lake Hancock are. In other words, fish appear to be less common in Lake Oklawaha than in Lake Hancock. The estimated odds ratio of exp(−1.1256) = 0.32 is the same for alligators of all sex and sizes, because this is a model with main effects but no interactions; see the entry "Lakeoklawaha:1" in the table from the R command **exp(coefficients(fit5))**. However, this coefficient is not statistically significant so these differences are not statistically significant because the Wald X^2=0.8912 (p-value=0.3452) which corresponds to the findings from the 95% CI for the odds-ratio estimates (0.031, 3.358) that contains 1; this CI can be computed as:

```
> exp(-1.1256+1.96*1.19230)
[1] 3.357874
> exp(-1.1256-1.96*1.19230)
[1] 0.03135103
```

```
> exp(coefficients(fit5))
 (Intercept):1  (Intercept):2  (Intercept):3  (Intercept):4   Lakegeorge:1   Lakegeorge:2   Lakegeorge:3   Lakegeorge:4 Lakeoklawaha:1 Lakeoklawaha:2 Lakeoklawaha:3
    0.08515564     0.12562535     0.05425079     0.39982590     0.56255501     5.93289534     0.32320675     0.46460153     0.32445217    14.78619873     4.05843171
Lakeoklawaha:4 Laketrafford:1 Laketrafford:2 Laketrafford:3 Laketrafford:4     Size>2.3:1     Size>2.3:2     Size>2.3:3     Size>2.3:4      Genderf:1      Genderf:2
    0.47686707     1.93813088    18.84663158     6.90044926     2.20601430     2.07557742     0.26282653     1.74549121     0.74782762     1.83387028     1.58877439
     Genderf:3      Genderf:4
    1.87303229     1.28732894
```

On the other hand, the estimated coefficient for the Lake Oklawaha dummy in the ***invertebrates-versus-fish*** equation is 2.6937 (st.error=0.6692, p-value 0.001) and highly significant. The estimated odds-ratio of 14.786 (with 95% CI [3.983, 54,893]) imply that alligators in Lake Oklawaha are more likely to choose invertebrates over fish than their colleagues in Lake Hancock are.

The above output also confirms that there is no significant *sex* effect and that size only matters for invertebrates versus fish food choice.

Goodness of fit

There are N = 16 profiles (unique combinations of lake, sex and size) in this dataset.

Recall that Residual deviance tests the fit of the current model versus the saturated model.

The saturated model, which fits a separate multinomial distribution to each profile, has 16 × 4 = 64 parameters.

The current model has an intercept, three lake coefficients, one sex coefficient and one size coefficient for each of the four logit equations, for a total of 24 parameters.

Therefore, the overall fit statistics have 64 − 24 = 40 degrees of freedom.

```
> pchisq(50.2637, df = 40, lower.tail = F)
[1] 0.1281848
> drop1(fit5, test = "LRT")
Single term deletions

Model:
cbind(Bird, Invertebrate, Reptile, Other, Fish) ~ Lake + Size +
    Gender
       Df Deviance    AIC    LRT  Pr(>Chi)
<none>      50.264 194.64
Lake   12  100.582 220.96 50.318 1.228e-06 ***
Size    4   67.864 204.24 17.600  0.001477 **
Gender  4   52.478 188.86  2.215  0.696321
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Does the model fit well?

The model fits OK but not great. The Residual Deviance of 50.26 with 40 df from the table above output is reasonable, with p-value of 0.1282 and the statistics/df is close to 1 that is 1.256. We will see that in terms of G^2/df and p-value, we can actually improve the fit by removing the gender covariate.

Model Selection

First, let's find the deviance G^2 for the ***null (intercept-only) model***, a model with just four parameters. Because there are $N = 4 \times 2 \times 2 = 16$ unique covariate patterns, the saturated model will have $16 \times (5 - 1) = 64$ parameters, so the G^2 statistic for the null model should have $64 - 4 = 60$ degrees of freedom.

```
> fit0

Call:
vglm(formula = cbind(Bird, Invertebrate, Reptile, Other, Fish) ~
    1, family = multinomial, data = gator)


Coefficients:
(Intercept):1 (Intercept):2 (Intercept):3 (Intercept):4
   -1.9783454    -0.4324209    -1.5988558    -1.0775589

Degrees of Freedom: 64 Total; 60 Residual
Residual deviance: 116.7611
Log-likelihood: -106.5708
```

Model	G^2	*df*
Saturated	0.00	0
Lake + Size + Lake × Size**	35.40	32
Lake + Size + Sex	50.26	40
Lake + Size	52.48	44
Lake	73.57	48
Size	101.61	56
Sex	114.66	56
Null	116.76	60
**Note: did not converge or issue with the fit		

```
> fit1=vglm(cbind(Bird,Invertebrate,Reptile,Other,Fish)~Lake * Size, data=gator, family=multinomial)
> deviance(fit1)
[1] 35.39866
> fit2=vglm(cbind(Bird,Invertebrate,Reptile,Other,Fish)~Lake + Size + Gender, data=gator, family=multinomial)
> deviance(fit2)
[1] 50.26369
> fit3=vglm(cbind(Bird,Invertebrate,Reptile,Other,Fish)~Lake + Size, data=gator, family=multinomial)
> deviance(fit3)
[1] 52.47849
> fit4 =vglm(cbind(Bird,Invertebrate,Reptile,Other,Fish)~Lake, data=gator, family=multinomial)
> deviance(fit4)
[1] 73.5659
> fit5 =vglm(cbind(Bird,Invertebrate,Reptile,Other,Fish)~Size, data=gator, family=multinomial)
> deviance(fit5)
[1] 101.6116
> fit6 =vglm(cbind(Bird,Invertebrate,Reptile,Other,Fish)~Gender, data=gator, family=multinomial)
> deviance(fit6)
[1] 114.6571
```

```
> drop1(fit2,test = "LRT")
Single term deletions

Model:
cbind(Bird, Invertebrate, Reptile, Other, Fish) ~ Lake + Size +
    Gender
       Df Deviance    AIC    LRT  Pr(>Chi)
<none>      50.264 194.64
Lake   12  100.582 220.96 50.318 1.228e-06 ***
Size    4   67.864 204.24 17.600  0.001477 **
Gender  4   52.478 188.86  2.215  0.696321
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

we can use the table above for model selection by comparing the models via calculation of ΔG^2, and Δdf. Our "final" model will have main effects for lake and size but no effect for sex (because "Lake+Size" - "Lake + Size +Sex" = 52.48 - 50.26 = 2.22 with 44 - 40 = 4 df which is not significant)

Based on the analysis-of-deviance table, our "final" model will have *main effects for lake and size but no effect for sex*

```
Call:
vglm(formula = cbind(Bird, Invertebrate, Reptile, Other, Fish) ~
    Lake + Size, family = multinomial, data = gator)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept):1   -2.0286     0.5581  -3.635 0.000278 ***
(Intercept):2   -1.7492     0.5392  -3.244 0.001178 **
(Intercept):3   -2.4230     0.6436  -3.765 0.000167 ***
(Intercept):4   -0.7465     0.3520  -2.121 0.033928 *
Lakegeorge:1    -0.6951     0.7813  -0.890 0.373608
Lakegeorge:2     1.6584     0.6129   2.706 0.006813 **
Lakegeorge:3    -1.2428     1.1854  -1.048 0.294461
Lakegeorge:4    -0.8262     0.5575  -1.482 0.138378
Lakeoklawaha:1  -1.3483     1.1633  -1.159 0.246453
Lakeoklawaha:2   2.5956     0.6597   3.934 8.34e-05 ***
Lakeoklawaha:3   1.2161     0.7860   1.547 0.121823
Lakeoklawaha:4  -0.8205     0.7296  -1.125 0.260753
Laketrafford:1   0.3926     0.7818   0.502 0.615487
Laketrafford:2   2.7803     0.6712   4.142 3.44e-05 ***
Laketrafford:3   1.6925     0.7804   2.169 0.030113 *
Laketrafford:4   0.6902     0.5597   1.233 0.217512
Size>2.3:1       0.6307     0.6425   0.982 0.326291
Size>2.3:2      -1.4582     0.3959  -3.683 0.000231 ***
Size>2.3:3       0.3513     0.5800   0.606 0.544785
Size>2.3:4      -0.3316     0.4483  -0.740 0.459511
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Names of linear predictors: log(mu[,1]/mu[,5]), log(mu[,2]/mu[,5]), log(mu[,3]/mu[,5]), log(mu[,4]/mu[,5])

Residual deviance: 52.4785 on 44 degrees of freedom

Log-likelihood: -74.4295 on 44 degrees of freedom

Number of Fisher scoring iterations: 5

Warning: Hauck-Donner effect detected in the following estimate(s):
'(Intercept):3'
```