Psy 5011: Categorical Data Analysis **(Due: 31/03/20)**

Homework #3

Instructor: Yao, Kaiping Grace *Name:* Hao-Cheng Lo, *Id:* D08227104

Problem 3.2: In the years 1904, 1914, 1924, . . . , 2014, the percentage of times the starting pitcher pitched a complete game were: 87.6, 55.0, 48.7, 43.4, 45.2, 34.0, 24.5, 28.0, 15.0, 8.0, 3.1, 2.4.

(a) The linear probability model has least squares fit $\hat{P}(Y = 1) = 0.6930 - 0.0662x$, where x = number of decades since 1904. Interpret 0.0662.

-0.0662 means that a fixed (i.e. 1) change in an x, the value of $\hat{P}(Y = 1)$ changes (i.e. the rate of change). In this case, for every decade, the percentage of times the starting pitcher pitched a complete game would be decrease expected 6.62 percentage, given linear prob. model.

(b) Substituting $x = 12$ in the linear prediction equation, predict the proportion of complete games for 2024. The ML fit of the logistic regression model yields $\hat{P}(Y = 1) = 0.034$ at $x = 12$. Which prediction is more plausible? Why?

(i) After substituting $x = 12$, we get $\hat{P}(Y = 1) = 0.6930 - 0.0662 \times 12 = -0.1014$.

(ii) Suppose the ML fit of the logistic regression model yields $\hat{P}(Y = 1) = 0.034$ at $x = 12$ is correct. The prediction of logistic regression model is more plausible. First, due to the characteristic of response data, the random component should be considered as binomial distributed where $0 \leq \pi \leq 1$ and link function should be considered as its canonical link *logistic*. Therefore, we can find that the prediction of logistic regression model at $x = 12$ locates at reasonable range $[0, 1]$ while linear model does not. Second, we can find that fixed change in an x may have less impact when $P(Y = 1)$ is near 0 or 1 than when $P(Y = 1)$ is near the middle of its range, which fit the logistic function intuitively. Hence, the prediction of logistic regression model is more plausible.

Problem 3.6: From the 2016 General Social Survey, when we cross-classify political ideology (with 1 being most liberal and 7 being most conservative) by political party affiliation for subjects of ages 18–27, we get:

```
            1   2   3   4   5   6   7
Democrat    5   18  19  25  7   7   2
Republican  1   3   1   11  10  11  1
```

When we use R to model the effect of political ideology on the probability of being a Democrat, we get the results:

```
> y <- c(5,18,19,25,7,7,2); n <- c(6,21,20,36,17,18,3)
> x <- c(1,2,3,4,5,6,7)
> fit <- glm(y/n ~ x, family=binomial(link=logit), weights=n)
> summary(fit)
#            Estimate    Std. Error    z value   Pr(>|z|)
#(Intercept)  3.1870        0.7002      4.552   5.33e-06
#x           -0.5901        0.1564     -3.772   0.000162
#---
#    Null deviance: 24.7983  on 6 degrees of freedom
#Residual deviance:  7.7894  on 5 degrees of freedom
#Number of Fisher Scoring iterations: 4
> confint(fit)
#                2.5 %       97.5 %
#(Intercept)  1.90180      4.66484
#          x  -0.91587     -0.29832
```

(a) Report the prediction equation and interpret the direction of the estimated effect.

The ML fit of the logistic regression model:

$$\text{logit}[\hat{\pi}(x)] = 3.1870 - 0.5901x$$

For the intercept, $\hat{\alpha} = 3.1870$ refers that the $f : x \to \hat{\pi}(x)$ shifts right more than others whose $\hat{\alpha}$ is smaller. For the estimate effect *magnitude*, $|\hat{\beta}| = 0.5901$ refers that the curve of f is steeper than others whose $|\hat{\beta}|$ is smaller. For the estimate effect *direction* , if $\hat{\beta} = -0.5901 < 0$ then $\hat{\pi}(x)$ decreases as x increases, which refers that in this case, as people more conservative, the prob. that he affiliates Democrat decreases.

(b) Construct the 95% Wald confidence interval for the effect of political ideology. Interpret and compare to the profile likelihood interval shown.

The Wald 95% confidence interval for β is $\hat{\beta} \pm 1.96(SE)$, which is $-0.5901 \pm 1.96(0.1564) = (-0.897, -0.284)$. That is, we have 95% confidence that β (the true value of log odds-ratio with one unit increase in x) would be in (-0.897, -0.284). From the R output, the profile likelihood 95% confidence interval for β is $(-0.916, -0.298)$. The Wald CI and profile likelihood CI are slightly different. If the distribution of the parameter estimator is skewed or sample size is small or moderate, Wald CI could perform poorly. Profile likelihood CI don't assume normality of the estimator, therefore perform better for small sample size than Wald CI.

(c) Conduct the Wald test for the effect of x. Report the test statistic, P-value, and interpret.

The Wald test of $H_0 : \beta = 0$ against $H_a : \beta \neq 0$ treats the test statistic as standard normal under H_0, which is:

$$z = \frac{\hat{\beta}}{SE} = \frac{-0.5901}{0.1564} = -3.773 \text{ or } z^2 = 14.24 \text{ as } \chi^2 \text{ with } df = 1$$

P-value = .00016, providing extremely strong evidence that political party affiliation as Democrat decreases as political ideology is more conservative.

(d) Conduct the likelihood-ratio test for the effect of x. Report the test statistic, find the P-value, and interpret.

For conducting the likelihood-ratio test, in R:

```
> library(car)
> Anova(fit)
#  LR Chisq Df Pr(>Chisq)
#x   17.009  1   3.72e-05 ***
#---
#Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The likelihood-ratio test of $H_0 : \beta = 0$ against $H_a : \beta \neq 0$ treats the test statistic as χ^2 with $df = 1$ under H_0, which is:

$$G^2 = 2(L_1 - L_0) = 17$$

P-value = .0000372, providing extremely strong evidence that political party affiliation as Democrat decreases as political ideology is more conservative.

(e) Explain the output about the number of Fisher scoring iterations.

Because ML estimates for GLM has no analytical form. Hence, we need a numerical algorithm, Fisher scoring algorithm, to find that. The Fisher scoring algorithm refers that started with given a set of random seed as parameters' values, then iteratively do *weighted least squares fitting*. From cycle to cycle, the weights change with revised approximations for the ML estimates and variance estimates. At the end, iteration terminates when ML estimates' difference is acceptable small (i.e. convergence). The number of Fisher scoring iterations in the output indicates that there do 4 cycles to converge and find the the ML estimates and variance estimates.

Problem 3.7: Consider Table 3.1 on snoring and heart disease.

Table 3.1 Relationship between snoring and heart disease, with model fits for the proportion of *yes* responses.

Snoring	Heart Disease		Proportion	Linear	Logistic
	Yes	No	Yes	Fit	Fit
Never	24	1355	0.017	0.017	0.021
Occasional	35	603	0.055	0.057	0.044
Nearly every night	21	192	0.099	0.096	0.093
Every night	30	224	0.118	0.116	0.132

Source of data: P.G. Norton and E.V. Dunn, *Brit. Med. J.* **291**: 630–632 (1985), published by BMJ Publishing Group.

(a) Re-fit the logistic regression model using the scores (i) (0, 2, 4, 6), (ii) (0, 1, 2, 3), (iii). (1, 2, 3, 4). Compare the model parameter estimates under the three choices. Compare the fitted values. What can you conclude about the effect of *linear* transformations of scores that preserve relative sizes of spacings between scores?

Using R to find the refit the scores of (i), (ii), (iii):

```
> Heart <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Heart.dat", header=TRUE)
> n <- Heart$yes + Heart$no
> x1 <- c(0,2,4,6)
> x2 <- c(0,1,2,3)
> x3 <- c(1,2,3,4)
> yes <- Heart$yes
> fit1 <- glm(yes/n ~ x1, family=binomial(link=logit), weights=n)
> fit2 <- glm(yes/n ~ x2, family=binomial(link=logit), weights=n)
> fit3 <- glm(yes/n ~ x3, family=binomial(link=logit), weights=n)
> summary(fit1)
#Coefficients: (the model parameter estimates of (i))
#            Estimate Std. Error z value Pr(>|z|)
#(Intercept) -3.77738    0.15571 -24.260  < 2e-16 ***
#x1           0.32726    0.04126   7.931 2.18e-15 ***
> summary(fit2)
#Coefficients: (the model parameter estimates of (ii))
#            Estimate Std. Error z value Pr(>|z|)
#(Intercept) -3.77738    0.15571 -24.260  < 2e-16 ***
#x2           0.65453    0.08253   7.931 2.18e-15 ***
> summary(fit3)
#Coefficients: (the model parameter estimates of (iii))
#            Estimate Std. Error z value Pr(>|z|)
#(Intercept) -4.43191    0.22557 -19.647  < 2e-16 ***
#x3           0.65453    0.08253   7.931 2.18e-15 ***
> fitted(fit1)
#         1          2          3          4
#0.02237076 0.04217465 0.07810938 0.14018107 (the fitted value of (i))
> fitted(fit2)
#         1          2          3          4
#0.02237076 0.04217465 0.07810938 0.14018107 (the fitted value of (ii))
> fitted(fit3)
#         1          2          3          4
#0.02237076 0.04217465 0.07810938 0.14018107 (the fitted value of (iii))
```

As for the model parameter estimates, we can find that from (i) to (ii) scores are scaled by 0.5 so that the $\hat{\beta}_0^{(ii)}$ remains (i.e. equals to $\hat{\beta}_0^{(i)}$), the $\hat{\beta}_1^{(ii)}$ equals $\hat{\beta}_1^{(i)}$ scaled by $0.5^{-1} = 2$; we can also find that from (ii) to (iii) scores are translated to right by 1 unit so that the $\hat{\beta}_1^{(iii)}$ remains (i.e. equals to $\hat{\beta}_1^{(ii)}$), the $\hat{\beta}_0^{(iii)}$ becomes $\hat{\beta}_0^{(ii)} + \hat{\beta}_1^{(ii)} \times (-1) \times 1$.

As for the fitted values, linear transformation of scores won't affect the fitted values.

To generally illustrate this phenomenon, take LSE for example (see pf (1)). Let A be the augmented data matrix, K be the linear transformation matrix, which is a square matrix, and y be the response data vector.

$$\begin{aligned}\hat{y}_{new} &= AK((AK)^T(AK))^{-1}(AK)^T y = AK(K^T(A^TA)K)^{-1}K^TA^Ty \\ &= AKK^{-1}(A^TA)^{-1}(K^T)^{-1}K^TA^Ty = A(A^TA)^{-1}A^Ty = \hat{y}_{old}\end{aligned} \tag{1}$$

We can find that for fitted value, the new one equals to the old one. And for the model parameter estimates, the new one is the old one multiplied by K^{-1}. From (i) to (ii), K is a scaling matrix, and from (ii) to (iii), K is a translation matrix.

(b) Fit the logistic regression model using the scores (0, 2, 6, 7), approximating the number of days in a week that the subject snores. Compare fitted values to those with the scores (0, 2, 4, 5) used in the text example. Do results seem to be sensitive to the choice of scores?

Using R to find the refit the scores of (0,2,6,7) and (0,2,4,5):

```
> x4 <- c(0,2,6,7)
> x5 <- c(0,2,4,5)
> fit4 <- glm(yes/n ~ x4, family=binomial(link=logit), weights=n)
> fit5 <- glm(yes/n ~ x5, family=binomial(link=logit), weights=n)
> fitted(fit4)
          1           2           3           4
#0.02288749 0.03807029 0.10151539 0.12805716
> fitted(fit5)
          1           2           3           4
#0.02050742 0.04429511 0.09305411 0.13243885
```

The fitted values of 2 sets of scores are different because (0,2,6,7) and (0,2,4,5) are not linear transformation. Given this situation, the results seem to be sensitive the choice of scores.

Problem 3.10: A recent General Social Survey asked "How many people at your work place are close friends?" The 756 responses had a mean of 2.76, standard deviation of 3.65, and a mode of 0. Would the Poisson distribution describe these data well? Why or why not?

We can found that the data is a count data with a right-tailed distribution where $\mathrm{VAR}(X) > \mathrm{E}(X)$. The Poisson distribution may not describe these data well, because there overdispersion occurs, which indicating discrete data over an unbounded positive range whose sample variance exceeds the sample mean. Hence, the observations are overdispersed voilating a Poisson distribution, where the mean is equal to the variance. Hence a Poisson distribution is not an appropriate model. The negative binomial is an appropriate model to describe the data, because it has one more parameter than the Poisson, the second parameter can be used to adjust the variance independently of the mean.

Problem 3.13: For the Crabs data file, fit the Poisson loglinear model to use weight to predict the number of satellites.

(a) Report the prediction equation, and estimate the mean response for female crabs of average weight, 2.44 kg.

Using R to test model:

```
> Crabs <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Crabs.dat", header=TRUE)
> fit <- glm(sat ~ weight, family=poisson(link=log), data=Crabs)
> summary(fit)
#Deviance Residuals:
#    Min       1Q   Median       3Q      Max
#-2.9307  -1.9981  -0.5627   0.9298   4.9992
#
#Coefficients:
#             Estimate Std. Error z value Pr(>|z|)
#(Intercept) -0.42841    0.17893  -2.394   0.0167 *
#weight       0.58930    0.06502   9.064   <2e-16 ***
#---
#Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1
#
#(Dispersion parameter for poisson family taken to be 1)
#
#    Null deviance: 632.79  on 172  degrees of freedom
#Residual deviance: 560.87  on 171  degrees of freedom
#AIC: 920.16
#
#Number of Fisher Scoring iterations: 5
> new <- data.frame(weight = c(2.44))
> predict.glm(fit, new, type = "resp")
#       1
#2.74422
```

To detect overdispersion:

```
> P__disp(fit)
# pearson.chi2   dispersion
#  535.895723     3.133893
> dispersiontest(fit)
#z = 5.3758, p-value = 3.812e-08
#alternative hypothesis: true dispersion is greater than 1
#sample estimates:
#dispersion
#  3.116407
```

The prediction equation is:

$$\log(\hat{\mu}) = -0.428 + 0.589(\text{weight})$$

The mean response of sat for female crabs of average weight $= 2.74$.

Overdispersion: we found that the data has overdispersion. Hence, we may do some corrections. A common approach to take into account over-dispersion in inference is to take ϕ into account for count data Y.

Estimation of ϕ using the *Pearson* statistic:

$$\hat{\phi}_P = 3.13$$

(b) Use $\hat{\beta}$ to describe the weight effect. Construct a 95% confidence interval for β and for the multiplicative effect of a 1-kg increase.

A one unit increase unit in weight has a multiplicative impact of $e^{\hat{\beta}}$ on $\hat{\mu}$. Because $\hat{\beta} > 0$, $e^{\hat{\beta}} > 1$, and as weight increases the sat of crabs increase.

For constructing CI:

Without Overdispersion:

The Wald 95% confidence interval for β is $\hat{\beta} \pm 1.96(SE)$, which is $0.589 \pm 1.96(0.065) = 0.589 \pm 0.127$ or $(0.462, 0.717)$. For this model, $e^{\hat{\beta}} = 1.80$ represents the multiplicative effect on the fitted value for each 1-kg increase in weight. The CI for the multiplicative effect on the mean is $(1.59, 2.05)$.

With Overdispersion:

The corrected Wald 95% confidence interval for β is $\hat{\beta} \pm 1.96(SE\sqrt{\hat{\phi}_P})$, which is $0.589 \pm 1.96(0.065 \times 1.77) = 0.589 \pm 0.225$ or $(0.364, 0.814)$. The CI for the multiplicative effect on the mean is $(1.44, 2.26)$.

(c) Conduct Wald and likelihood-ratio tests of the hypothesis that the mean response is independent of weight. Interpret.

Without Overdispersion:

The Wald test of $H_0 : \beta = 0$ against $H_a : \beta \neq 0$ treats the test statistic as standard normal under H_0, which is:

$$z = \frac{\hat{\beta}}{SE} = \frac{0.589}{0.065} = 9.06 \text{ or } z^2 = 82.2 \text{ as } \chi^2 \text{ with } df = 1$$

P-value $< .0001$, providing extremely strong evidence that weight has a positive effect.

The likelihood-ratio test of $H_0 : \beta = 0$ against $H_a : \beta \neq 0$ treats the test statistic as χ^2 with $df = 1$ under H_0, which is:

$$G^2 = 2(L_1 - L_0) = 632.79 - 560.87 = 71.9$$

P-value $< .0001$, providing extremely strong evidence that weight has a positive effect.

With Overdispersion:

The Wald test of $H_0 : \beta = 0$ against $H_a : \beta \neq 0$ treats the test statistic as standard normal under H_0, which is:

$$z = \frac{\hat{\beta}}{SE\sqrt{\hat{\phi}_P}} = \frac{0.589}{0.065 \times 1.77} = 5.12 \text{ or } z^2 = 26.21 \text{ as } \chi^2 \text{ with } df = 1$$

P-value $< .0001$, providing extremely strong evidence that weight has a positive effect.

The likelihood-ratio test of $H_0 : \beta = 0$ against $H_a : \beta \neq 0$ treats the test statistic as χ^2 with $df = 1$ under H_0, which is:

$$G^2 = \frac{2(L_1 - L_0)}{\hat{\phi}_P} = 71.9/3.13 = 22.97$$

P-value $< .0001$, providing extremely strong evidence that weight has a positive effect.