Psy 5011: Categorical Data Analysis **(Due: 14/04/20)**

Homework #4

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

Problem 4.1: A study investigated characteristics associated with y = whether a cancer patient achieved remission (1 = yes, 0 = no). An important explanatory variable was a labeling index (LI = percentage of "labeled" cells) that measures proliferative activity of cells after a patient receives an injection of tritiated thymidine. R code shows the data and R output for a logistic regression model.

```
> LI <- c(8,8,10,10,12,12,12,14,14,14,16,16,16,18,20,20,20,22,22,24,26,28,32,34,38,38,38)
> y <- c(0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,0,1,0,0,1,1,0,1,1,1,0)
> summary(glm(y ~ LI, family=binomial))
#              Estimate    Std. Error  z value Pr(>|z|)
#(Intercept)    -3.77714    1.37862     -2.740  0.00615
#LI             0.14486     0.05934     2.441   0.01464
#---
#    Null deviance: 34.372 on 26 degrees of freedom
#Residual deviance: 26.073 on 25 degrees of freedom
> confint(glm(y ~ LI, family=binomial))
#   2.5 %   97.5 %
#LI 0.04252 0.28467
```

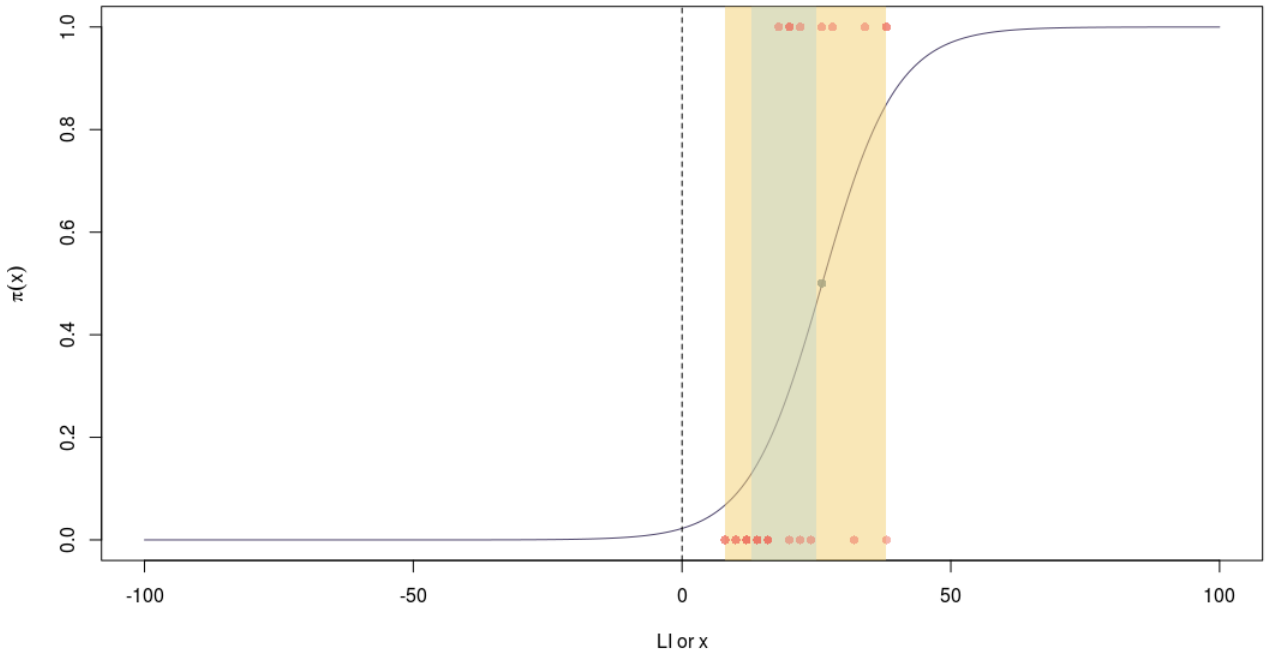
(a) Show that $\hat{P}(Y = 1) = 0.50$ when $LI = 26.0$.

When $LI = 26.0$, $\hat{P}(Y = 1) = \dfrac{e^{-3.7771+0.1449\times 26}}{1 + e^{-3.7771+0.1449\times 26}} = 0.5$.

(b) When LI increases by 1, show that the estimated odds of remission multiply by 1.16.

$\hat{\text{Odds}}_{x+1} = e^{\hat{\alpha}+\hat{\beta}(x+1)} = e^{\hat{\alpha}+\hat{\beta}(x)}e^{\hat{\beta}} = \hat{\text{Odds}}_x e^{\hat{\beta}} = \hat{\text{Odds}}_x e^{0.1449} = 1.16\hat{\text{Odds}}_x$.

(c) Summarize the LI effect by how $\hat{P}(Y = 1)$ changes over the range or interquartile range of LI values.



(c) *Conti.*

To illustrate how $\hat{P}(Y = 1)$ (also donated as $\pi(x)$) changes over the range of LI values (also donated as x). I plotted the *estimated model* (dark curve), *raw data* (coral dots), *range of raw data* (yellow region), and *IQR of raw data* (steel-blue region) by R as shown above. There are following facts: First, for the estimate effect *direction*, if $\hat{\beta} = 0.1449 > 0$ then $\pi(x)$ increases as LI increases. Second, for the estimate effect *magnitude*, $|\hat{\beta}| = 0.1449$ relates to how steep the curve is, because the rate of change is $\pi'(x) = \hat{\beta}\pi(x)(1 - \pi(x))$. We can find that when $LI \to \pm\infty$, fixed change in an LI may have less impact on $\pi(x)$ than when LI is near the middle of its range. Note that the rate of change is maximized when $\pi(x) = 0.5$, implying max rate of change is $\hat{\beta}/4 = 0.0362$ when $x = -\alpha/\beta = 26.0$ (see the green point in the figure). It is noteworthy that the domain of LI is [0,100], which is the region right to the vertical dash line.

```
> predict(fit1,data.frame(LI=quantile(LI)), type="resp")
#         0%        25%        50%        75%       100%
#0.06797405 0.13079831 0.23692681 0.46118815 0.84911301
```

Note that the change in estimated prob. of range is $.85 - .07 = .78$ and that of IQR is $.46 - .13 = .33$.

(d) Show that the rate of change in $\hat{P}(Y = 1)$ is 0.009 when $LI = 8$.

When $LI = 8$, $\hat{P}(Y = 1) = .068$. Hence, rate of change is $\hat{\beta}\hat{\pi}(x)(1 - \hat{\pi}(x)) = 0.1449(0.068)(0.932) = 0.009$.

(e) Summarize the LI effect by the estimated average marginal effect.

```
> logitmfx(fit1,data.frame(LI = LI, y = y),atmean=FALSE)
#
#Marginal Effects:
#      dF/dx Std. Err.     z P>|z|
#LI 0.022584  0.014722 1.534 0.125
```

The estimated average marginal effect is the effect averaging the rate of change at the n sample values of the explanatory variables. We can calculate it by $\hat{\beta}\hat{\mathbf{y}}^T(1 - \hat{\mathbf{y}})/n = .023$. That is, at the $n = 27$ observed LI values, the average rate of change is 0.023 in the estimated probability of achieving remission per 1-percent increase in LI.

Problem 4.2: Refer to the previous exercise.

(a) Conduct a Wald test for the LI effect and construct a 95% Wald confidence interval for the odds ratio corresponding to a 1-unit increase in LI. 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:

$$z = \frac{\hat{\beta}}{SE} = \frac{0.14486}{0.05934} = 2.44 \text{ or } z^2 = 5.97 \text{ as } \chi^2 \text{ with } df = 1$$

P-value = .0146, providing extremely strong evidence that LI has a positive effect on remission. The 95% Wald confidence interval of β: $[0.1449 \pm 1.96 \times .0593] = [.03, .26]$, and the 95% Wald CI for e^{β}: $[1.03, 1.30]$. Therefore, we have 95% confidence that the odds of remission at $LI = x + 1$ are estimated in between 1.03 and 1.30 times the odds of remission at $LI = x$.

(b) Conduct a likelihood-ratio test and construct a 95% profile likelihood interval. Interpret.

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:

$$G^2 = 2(L_1 - L_0) = 34.372 - 26.073 = 8.3$$

P-value = .004, providing extremely strong evidence that LI has a positive effect on remission. The 95% profile likelihood interval of β: $[0.042, 0.284]$, which means we have 95% confidence that true LI effect would be in between 0.042 and 0.284. The profile likelihood interval of e^{β}: $[1.04, 1.33]$. Therefore, we have 95% confidence that the odds of remission at $LI = x + 1$ are estimated in between 1.04 and 1.33 times the odds of remission at $LI = x$.

Problem 4.4: For the snoring and heart disease data of Table 3.1 with snoring-level scores (0, 2, 4, 5), the logistic regression ML fit is $[\hat{P}(Y = 1)] = -3.866 + 0.397x$. Interpret the effect of snoring on the odds of heart disease.

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

	Heart Disease		Proportion	Linear	Logistic
Snoring	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.

To interpret the effect of snoring on the odds of heart disease. We can find that the multiplicative effect on odds of heart disease is $e^{\hat{\beta}} = e^{0.397} = 1.49$. Hence, 1.49 for one-unit change in snoring, and $1.49^2 = 2.21$ for two-unit change in snoring. For example, when snoring-level from *never* (0) to *occasional* (2), the odds of heart disease at *occasional* are estimated 2.21 times the odds of heart disease at *never*. Namely, the odds of heart disease is increased by 121% from *never* to *occasional*. Similarly, when snoring-level from *nearly every night* (4) to *every night* (5), the odds of heart disease at *every night* are estimated 1.49 times the odds of heart disease at *nearly every night*. Namely, the odds of heart disease is increased by 49% from *nearly every night* to *every night*.

Problem 4.9: For the Crabs data file, fit a logistic regression model for the probability of a satellite, using color alone as the predictor.

```
> crabs <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Crabs.dat",header=T)
> fitqt <- glm(y ~ color,
+              family=binomial(link=logit),
+              data=crabs)
> crabs$color <- factor(crabs$color)
> fitql <- glm(y ~ color,
+             family=binomial(link=logit),
+             contrasts=list(color=contr.treatment(4,base=4,contrasts=TRUE)),
+             data=crabs)
> summary(fitql)
#Coefficients:
#             Estimate Std. Error z value Pr(>|z|)
#(Intercept)  -0.7621     0.4577  -1.665 0.095910 .
#color1        1.8608     0.8087   2.301 0.021393 *
#color2        1.7382     0.5123   3.393 0.000692 ***
#color3        1.1299     0.5509   2.051 0.040289 *
#
#    Null deviance: 225.76  on 172  degrees of freedom
#Residual deviance: 212.06  on 169  degrees of freedom
#AIC: 220.06
> summary(fitqt)
#Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.3635     0.5551   4.257 2.07e-05 ***
color        -0.7147     0.2095  -3.412 0.000645 ***
#
#    Null deviance: 225.76  on 172  degrees of freedom
#Residual deviance: 213.30  on 171  degrees of freedom
#AIC: 217.3
```

(a) Treat color as a nominal-scale factor. Report the prediction equation and explain how to use it to compare the first and fourth colors.

For the qualitative *color c* variable, there are four levels $\in \{1, 2, 3, 4\}$. Consider the model for $\pi(c) = P[Y = 1|c1, c2, c3, c4]$, We have the following prediction equation:

$$\pi(c) = \alpha + \beta_1 c_1 + \beta_2 c_2 + \beta_3 c_3$$

Where c_1 is dummy variable for *color 1* ($c_1 = 1$ for *color 1*, 0 otherwise), c_2 is dummy variable for *color 2* ($c_2 = 1$ for *color 2*, 0 otherwise), c_3 is dummy variable for *color 3* ($c_3 = 1$ for *color 3*, 0 otherwise). Note that *color 4* is used as a reference color (when $c_1 = c_2 = c_3 = 0$). In order to compare the first and fourth colors, we can inspect β_1, which is the log odds-ratio of having at least one satellite between *color 1* crabs and *color 4* crabs. Namely, the effect of the *color 1* crabs compared to *color 4* on having at least one satellite.

Note that we can find that the fitted model is $\pi(c) = -0.76 + 1.86c_1 + 1.74c_2 + 1.13c_3$, and the odds that *color 1* crabs have at least one satellite is $e^{\hat{\beta}_1} = e^{1.86} = 6.42$ times the odds that *color 4* crabs have at least one satellite.

(b) For the model in (**a**), conduct a likelihood-ratio test of the hypothesis that color has no effect. Interpret.

The likelihood-ratio test of $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ against $H_a : \exists \beta_i \neq 0,\ i \in \{1, 2, 3\}$ treats the test statistic as χ^2 with $df = 3$ under H_0, which is:

$$G^2 = 2(L_1 - L_0) = 225.76 - 212.06 = 13.7$$

P-value = .003343, providing evidence that at least one kind of colors has effect on having satellite.

(c) Treating color in a quantitative manner (scores 1, 2, 3, 4), obtain a prediction equation. Interpret the coefficient of color and test the hypothesis that color has no effect.

For the quantitative *color c* variable, consider the model for $\pi(c) = P[Y = 1|c]$, We have the following prediction equation:

$$\pi(c) = \alpha + \beta c$$

The fitted model is $\pi(c) = 2.36 - 0.72c$. The coefficient of color is $\hat{\beta} = -0.72$ which is log odds-ratio of crab having at least one satellite with 1 unit increase of color (i.e. getting one unit darker). We can also find that the multiplicative effect on odds of crab having at least one satellite is $e^{\hat{\beta}} = e^{-0.72} = 0.48$. Hence, the odds of having at least one satellite at color $= x+1$ are estimated 0.48 times the odds of having at least one satellite at color $= x$.

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:

$$G^2 = 2(L_1 - L_0) = 225.76 - 213.30 = 12.46$$

P-value = .000416, providing evidence that colors has effect on having satellite.

(d) When we treat color as quantitative instead of qualitative, state a potential advantage relating to power and a potential disadvantage relating to model lack of fit.

As to the predictive power, generally speaking, tests of the effect of the variable are generally more powerful when it has a single parameter rather than several parameters. Hence, the quantitative treatment with $df = 1$ and P-value = .0004 potentially has more power than the qualitative treatment with $df = 3$ and P-value = .0033. As to the model fit, by observing the AIC, the AIC of quantitative treatment (= 217) is less than the AIC of qualitative treatment (= 220), indicating the former one has relative bad quality of statistical models for a given set of data than the later one. Hence, quantitative treatment has a potential disadvantage relating to model lack of fit.

Supplementary Information:

Use more sophisticated ways to inspect predictive power, including classification tables and ROC.

```
# Classification tables Method
> pihat.qt <- vector(length=173)
> for (i in 1:173){
+   pihat.qt[i] <-  predict.glm(update(fitqt, subset=-i),
+                               newdata=crabs[i,], type="response")
+ }
> yhat.qt <- as.numeric(pihat.qt > 0.50)
> confusionqt <- table(crabs$y, yhat.qt)
> confusionqt
#   yhat.qt
#      0   1
#  0  15  47
#  1   7 104
> pihat.ql <- vector(length=173)
> for (i in 1:173){
+   pihat.ql[i] <-  predict.glm(update(fitql, subset=-i),
+                                 newdata=crabs[i,], type="response")
+ }
> yhat.ql <- as.numeric(pihat.ql > 0.50)
> confusionql <- table(crabs$y, yhat.ql)
> confusionql
#   yhat.ql
#      0   1
#  0  15  47
#  1   7 104
```

(d) *Conti.*

```
# ROC method
> rocplotqt <- roc(y ~ fitted(fitqt), data=crabs)
> rocplotql <- roc(y ~ fitted(fitql), data=crabs)
> auc(rocplotqt)
#Area under the curve: 0.6386
> auc(rocplotql)
#Area under the curve: 0.6386
```

Based on above two methods, for color variable, the difference of the power of qualitative treatment and the power of quantitative treatment is very little.

For the model fit, I did a likelihood ratio test comparing two models.

```
> anova(fitql, fitqt, test="LRT")
#  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
#1       169     212.06
#2       171     213.30 -2   -1.237   0.5387
```

There is no significant difference between two models fits.

(e) Using weight and quantitative color as explanatory variables, find standardized coefficients, and interpret.

```
> fit2 <- glm(y ~ color+weight,
+             family=binomial(link=logit),
+             data=crabs)
> beta(fit2)
#Coefficients:
#            Estimate Std. Error z value Pr(>|z|)
#(Intercept)   0.7430     0.1851   4.013 5.98e-05 ***
#color.z      -0.4123     0.1791  -2.302   0.0213 *
#weight.z      0.9539     0.2207   4.322 1.55e-05 ***
#
#    Null deviance: 225.76  on 172  degrees of freedom
#Residual deviance: 190.27  on 170  degrees of freedom
#AIC: 196.27
```

Because a one-unit change in the standardized variable is a standard deviation change in the original variable. Then, $\hat{\beta}_j$ represents the effect of a standard deviation change in x_j, adjusting for the other variables. To illustrate, for standardized explanatory variables, the estimated effects equal 0.95 for weight and -0.41 for color. A standard deviation increase in weight is estimated to have more than double the effect of a standard deviation increase in color, adjusting for the other variable. Weight has a positive effect on having at least one satellite, while color has a negative effect.

Problem 4.17: Table 4.11 shows estimated effects for a logistic regression model for y = presence of squamous cell esophageal cancer (1 = yes, 0 = no). Smoking status (s) equals 1 for at least one pack per day and 0 otherwise, alcohol consumption (a) equals the average number of alcoholic drinks consumed per day, and race (r) equals 1 for blacks and 0 for whites.

Table 4.11 Table for Exercise 4.17 on effects on esophageal cancer.

Variable	Effect	P-value
Intercept	−7.00	<0.01
Alcohol use	0.10	0.03
Smoking	1.20	<0.01
Race	0.30	0.02
Race × Smoking	0.20	0.04

(a) To describe the race-by-smoking interaction, construct the prediction equation when $r = 1$ and again when $r = 0$. Find the fitted conditional odds ratio for the smoking effect for each case. Similarly, construct the prediction equation when $s = 1$ and again when $s = 0$. Find the fitted conditional odds ratio for the race effect for each case. (For each association, the coefficient of the cross-product term is the difference between the log odds ratios at the two levels for the other variable.)

The prediction equation is $[\hat{P}(Y = 1)] = -7 + 0.1a + 1.2s + 0.3r - 0.2(rs)$.

For $r = 1$:

$$[\hat{P}(Y = 1)] = -7 + 0.1a + 1.2s + 0.3 \times 1 - 0.2(1 \times s) = -6.7 + 0.1a + s$$

The ML estimated effects is $\hat{\beta}_{s|r=1} = 1$ for smoking. The estimated conditional odds ratio between cancer and smoking equals $e^1 = 2.71$. Hence, for each case, given the race is black ($r = 1$), the estimated odds that person who smoked ($s = 1$) had cancer were 2.71 times the estimated odds that person who didn't smoke ($s = 0$) had cancer, adjusting for alcohol use variable.

For $r = 0$:

$$[\hat{P}(Y = 1)] = -7 + 0.1a + 1.2s + 0.3 \times 0 - 0.2(0 \times s) = -7 + 0.1a + 1.2s$$

The ML estimated effects is $\hat{\beta}_{s|r=0} = 1.2$ for smoking. The estimated conditional odds ratio between cancer and smoking equals $e^{1.2} = 3.32$. Hence, for each case, given the race is white ($r = 0$), the estimated odds that person who smoked ($s = 1$) had cancer were 3.32 times the estimated odds that person who didn't smoke ($s = 0$) had cancer, adjusting for alcohol use variable.

For $s = 1$:

$$[\hat{P}(Y = 1)] = -7 + 0.1a + 1.2 \times 1 + 0.3r - 0.2(r \times 1) = -5.8 + 0.1a + 0.1r$$

The ML estimated effects is $\hat{\beta}_{r|s=1} = 0.1$ for race. The estimated conditional odds ratio between cancer and race equals $e^{0.1} = 1.11$. Hence, for each case, given that person who smoked ($s = 1$), the estimated odds that black ($r = 1$) had cancer were 1.11 times the estimated odds that white ($r = 0$) had cancer, adjusting for alcohol use variable.

For $s = 0$:

$$[\hat{P}(Y = 1)] = -7 + 0.1a + 1.2 \times 0 + 0.3r - 0.2(r \times 0) = -7 + 0.1a + 0.3r$$

The ML estimated effects is $\hat{\beta}_{r|s=0} = 0.3$ for race. The estimated conditional odds ratio between cancer and race equals $e^{0.3} = 1.34$. Hence, for each case, given that person who didn't smoke ($s = 0$), the estimated odds that black ($r = 1$) had cancer were 1.34 times the estimated odds that white ($r = 0$) had cancer, adjusting for alcohol use variable.

(b) In Table 4.11, what do the coefficients of smoking and race represent? What hypotheses do their P-values refer to?

Recall the equation: $[\hat{P}(Y=1)] = \alpha + \beta_a a + \beta_s s + \beta_r r - \beta_{r\times s}(rs)$

For smoking effect, we can rewrite the equation as $[\hat{P}(Y=1)] = -7 + 0.1a + (1.2 - 0.2r)s + 0.3r$. Hence, the $1.2 - 0.2r = \text{logit}\{\pi(s+1), r\} - \text{logit}\{\pi(s,r)\}$, adjusting other variables. We can find that s effect (log odds-ratio) on cancer depends on r, that is, r is an effect modifier. Among the coefficients relating to s, $\hat{\beta}_s = 1.2$ is the base/common s effect (log odds ratio) on cancer which will not be affected by r, adjusting other variables. The hypothesis related to s is, given other variables fixed, $H_0 : \beta_s = 0$ against $H_a : \beta_s \neq 0$.

For race effect, we can rewrite the equation as $[\hat{P}(Y=1)] = -7 + 0.1a + 1.2s + (0.3 - 0.2)r$. Hence, the $0.3 - 0.2r = \text{logit}\{\pi(r+1), s\} - \text{logit}\{\pi(r,s)\}$, adjusting other variables. We can find that r effect (log odds-ratio) on cancer depends on s, that is, s is an effect modifier. Among the coefficients relating to r, $\hat{\beta}_r = 0.3$ is the base/common r effect (log odds ratio) on cancer which will not be affected by s, adjusting other variables. The hypothesis related to r is, given other variables fixed, $H_0 : \beta_r = 0$ against $H_a : \beta_r \neq 0$.