Biostat 513: **Exercise Set #3** DUE: April 29, 2002

Reading: Kleinbaum *Logistic Regression* Chapters 1 & 2

NOTE: Unless explicitly stated, direct computer output is **not** desired. Typically only part of the computer output is asked for (such as a confidence interval) and then proper interpretation of the statistics is requested.

DATA: The data for these exercises can be found on the class web page:
`http://courses.washington.edu/b513/` in the *Homework* directory (ie. click on Homework from the main Biostat 513 page).

Stratified Tables

1. Let's continue the analysis of the data collected by Tuyns et al. (1977) (`NewTuyns.dat`). Recall that the goal of the study is to characterize the cancer risk associated with both alcohol and tobacco consumption. In many applications `age` is a potential confounder. In class we looked at case/control status (Y) and AGE (in the file `NewTuyns.do` for HW#2 – output is on the web page for HW#3) and found an association between AGE and disease.

For this exercise we will look at some dichotomizations of TOB and ALC. Define:
NEWALC = 1 if alcohol consumption is $\geq$ 80g/day, 0 otherwise. Define
NEWTOB = 1 if tobacco consumption is $\geq$ 10g/day, 0 otherwise.
(see web page for STATA code).

(a) Evaluate whether AGE is associated with NEWTOB and with NEWALC. Summarize the relationship that you see by summarizing a few summary statistics from the tabulation if AGE and NEWTOB, and AGE and NEWALC.

(b) Analyze the relationship between Y and NEWALC by creating a 2×2 table. Quote and interpret the odds ratio estimate and a 95% CI for the crude odds ratio.

(c) Now adjust for AGE as a potential confounder. Quote and interpret an adjusted odds ratio and 95% CI for NEWALC (common OR obtained from M-H). State in simple terms how the meaning of this estimate is different from that obtained in (b). Is a common odds ratio plausible (support your answer)?

(d) Analyze the relationship between Y and NEWTOB by creating a 2×2 table. Quote and interpret the odds ratio estimate and a 95% CI for the crude odds ratio.

(e) Now adjust for AGE as a potential confounder. Quote and interpret an adjusted odds

ratio and 95% CI for NEWTOB (common OR obtained from M-H). State in simple terms how the meaning of this estimate is different from that obtained in (d). Is a common odds ratio plausible (support your answer)?

(f) Consider whether ALC and TOB are related. Summarize simple statistics that describe the magnitude of association (or lack of) between these variables.

(g) Perform a stratified analysis of NEWALC after stratification on both AGE and TOB. Quote and interpret an adjusted odds ratio and 95% CI for NEWALC (common OR obtained from M-H). State in simple terms how the meaning of this estimate is different from that obtained in (b) and (c). Is a common odds ratio plausible (support your answer)?

(h) Perform a stratified analysis of NEWTOB after stratification on both AGE and ALC. Quote and interpret an adjusted odds ratio and 95% CI for NEWALC (common OR obtained from M-H). State in simple terms how the meaning of this estimate is different from that obtained in (d) and (e). Is a common odds ratio plausible (support your answer)?

Logistic Regression

2. A study was undertaken to assess predictors of a baby being born with a low birth weight. A total of $n = 189$ women receiving care at Baystate Medical Center were surveyed and asked about their behavior during pregnancy (including diet, smoking, and prenatal care visits). Babies that are born weighing less than $2500g$ are considered "low birth weight" (LBW=1 if $<2500g$, 0 otherwise) and are at increased risk of morbidity and mortality. Documentation for the analysis variables is in the `Datasets` section of the course web. Briefly, SMOKE=1 if the mother is a smoker, AnyPTL=1 if any history of premature labor (ie. PTL$\geq$1), RACE(2)=1 if mother is Black and 0 otherwise, RACE(3)=1 if mother is neither White nor Black (ie. RACE="other") and 0 otherwise, HYPER=1 if mother has history of hypertension and 0 otherwise, URIRR=1 if presence of uterine irritability and 0 otherwise, AGE20 = maternal age - 20 (years).

In analyzing the data, two logistic models were fit, each involving the dependent variable LBW, but with different sets of independent variables. The variables involved in each model and their estimated coefficients are listed below:

Model 1		Model 2	
Variable	Coefficient	Variable	Coefficient
Intercept	-2.003	Intercept	-1.905
SMOKE	0.896	SMOKE	0.713
AnyPTL	1.317	AnyPTL	1.295
RACE(2)	0.962	RACE(2)	0.870
RACE(3)	0.951	RACE(3)	0.906
HYPER	1.364	HYPER	1.396
URIRR	0.769	URIRR	0.838
AGE20	-0.051	AGE20	-0.082
		SMOKE×AGE20	0.066

(a) For model 1 above, state the form of the logistic model that was used – stating the dependent variable, the interpretation of the probability $\pi(X)$, and the model for $\pi(X)$ in terms of the (unknown) population parameters and the independent variables.

(b) For model 1 in (a) state the form of the estimated log odds functions: $\text{logit}[\pi(X)] =$.

(c) Using model 1, compute the estimated risk for LBW (ie. P[LBW=1]) for a smoker (SMOKE=1), without a history of premature labor (AnyPTL=0), white (RACE(2)=0,RACE(3)=0), with hypertension (HYPER=1), without uterine irritability (URIRR=0), who is age 30 (person 1), and a non-smoker (SMOKE=0), without a history of premature labor (AnyPTL=0), white (RACE(2)=0,RACE(3)=0), with hypertension (HYPER=1), without uterine irritability (URIRR=0), who is age 30 (person 2), What is the estimated relative risk for these individuals?

(d) Repeat part (c) using model 2. Why is the estimate different?

(e) What is the estimated odds ratio comparing SMOKER=1 to SMOKER=0 for 20 year old women with AnyPTL=0, RACE(2)=0, RACE(3)=0, HYPER=0, and URIRR=0 under model 1 and under model 2 (Note: use the coefficients directly rather than calculate $\widehat{\pi}(X)$).

(f) What is the estimated odds ratio comparing SMOKER=1 to SMOKER=0 for for 30 year old women with AnyPTL=0, RACE(2)=0, RACE(3)=0, HYPER=0, and URIRR=0 under model 1 and under model 2 (Note: use the coefficients directly rather than calculate $\widehat{\pi}(X)$).

3. Logistic regression can be used to compute odds ratio estimates after adjusting for other variables. Consider the adjusted analyses performed in question 1 that focused on NEWALC as the exposure variable of interest.

(a) What would be the dependent variable in a logistic regression for the `NewTuyns` data? Use this to define $\pi(X)$ for this example.

(b) Define a logistic regression model that would characterize the unadjusted odds ratio that was estimated in question 1(b).

(c) The following output is from a logistic regression of Y on both NEWALC and AGE, using dummy variables for age categories. Compute and interpret the estimated odds ratio comparing NEWALC=1 to NEWALC=0 for individuals with the same age. Compare this estimate to that obtained in question 1(e). Is it of similar magnitude? Does it have a similar interpretation?

(d) The following output is from a logistic regression of Y on both NEWALC, AGE, and TOB, using dummy variables for age categories and tobacco categories. Compute and interpret the estimated odds ratio comparing NEWALC=1 to NEWALC=0 for individuals with the same age and the same tobacco consumption. Compare this estimate to that obtained in question 1(g). Is it of similar magnitude? Does it have a similar interpretation?

Note: In the logistic regression analysis the first two age groups have been combined as the reference group by using the variable `Nage`. This variable takes the values 2, 3, 4, 5, and 6, and subjects with `age`=1 have been recoded as `Nage`=2, while other categories remain unchanged:

```
. table Nage y

----------------------------
          |   Case/Control
          |      Status
     Nage | Control     Case
----------+-----------------
    25-44 |     302       10
    45-54 |     167       46
    55-64 |     166       76
    65-74 |     106       55
      75+ |      31       13
----------------------------
```

*** Regression Output on Next Page ***

NEWALC
======

Logit estimates — Number of obs = 972; LR chi2(5) = 196.28; Prob > chi2 = 0.0000; Log likelihood = -395.91449; Pseudo R2 = 0.1986

y	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
newalc	1.679329	.1896687	8.85	0.000	1.307585	2.051073
_INage_3	1.965179	.3705434	5.30	0.000	1.238927	2.69143
_INage_4	2.480356	.3579642	6.93	0.000	1.778759	3.181953
_INage_5	2.734736	.3714681	7.36	0.000	2.006672	3.4628
_INage_6	2.731282	.4757757	5.74	0.000	1.798779	3.663785
_cons	-3.824719	.3341435	-11.45	0.000	-4.479629	-3.16981

NEWALC
======

Logit estimates — Number of obs = 972; LR chi2(8) = 223.10; Prob > chi2 = 0.0000; Log likelihood = -382.50332; Pseudo R2 = 0.2258

y	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
newalc	1.624142	.1947106	8.34	0.000	1.242516	2.005768
_INage_3	2.031365	.3782274	5.37	0.000	1.290053	2.772678
_INage_4	2.575207	.3662371	7.03	0.000	1.857396	3.293019
_INage_5	2.981799	.3841714	7.76	0.000	2.228837	3.734761
_INage_6	2.877458	.4852636	5.93	0.000	1.926359	3.828558
_Itob_2	.4380761	.219562	2.00	0.046	.0077425	.8684097
_Itob_3	.69607	.2633902	2.64	0.008	.1798346	1.212305
_Itob_4	1.569025	.317935	4.94	0.000	.9458841	2.192167
_cons	-4.300036	.362274	-11.87	0.000	-5.010081	-3.589992