INSTRUCTIONS:
Complete with legible handwriting, or use a mathematical editor (like MS Word, \LaTeX). Combination
of the two is also OK. Agresti refers to the textbook [3rd edition].

DUE DATE: April 19, 2018

1. Agresti 5.19

2. Agresti 5.20

3. Agresti 5.24. I posted the article in the course webpage. Use stepwise selection with default
SLENTRY=SLSTAY=0.05 to arrive at a final model. Start from a model with all three main
effects and all three two-way interactions. Report the H-L GOF p-value. How about a plot of the
$r_i$ vs. $\hat{\eta}_i$ for $i = 1, \ldots, 8$, with a loess smooth?

```sas
data colds;
input colds total titer$ virus$ social$;
datalines;
25 33 '<=2' 'RV39' '1-5'
20 38 '<=2' 'RV39' '>6'
18 30 '<=2' 'Hanks' '1-5'
21 43 '<=2' 'Hanks' '>6'
11 34 '>4' 'RV39' '1-5'
8 42 '>4' 'RV39' '>6'
3 26 '>4' 'Hanks' '1-5'
3 30 '>4' 'Hanks' '>6'
;
```

4. Agresti 5.25

5. Agresti 5.26

6. Finney (1941) and Pregibon (1981) present data from a controlled study of the effect of the rate
and volume of air inspired (after a single quick breath) on whether a transient vasoconstriction
occurred in finger skin (0=no, 1=yes). The data are in `vaso.sas` [see course website]. Rate is in
liters/second and volume is in liters. From considerations in Finney (1941), the natural logarithm
of volume, log(volume), and the natural logarithm of rate, log(rate), will be used as main effects
– you will need to take the log in the data step.

(a) Fit binary regression models with logit, probit, and complimentary log-log links. The logistic
regression model is

$$\text{logit}\{P(V = 1)\} = \beta_0 + \beta_1 \log(\text{volume}) + \beta_2 \log(\text{rate}),$$

the complimentary log-log model is

$$P(V = 1) = 1 - \exp\{-\exp\{\beta_0 + \beta_1 \log(\text{volume}) + \beta_2 \log(\text{rate})\}\},$$

and the probit regression model is

$$P(V = 1) = \Phi\{\beta_0 + \beta_1 \log(\text{volume}) + \beta_2 \log(\text{rate})\}.$$
(b) Write down and simplify the probability of vasoconstriction from the fitted model. How does increasing rate or volume affect the probability of vasoconstriction?

(c) What does the Hosmer and Lemeshow test say about this model?

(d) Are there any ill-fit observations according to the Pearson residuals?

(e) Are there any influential observations in terms of the DFBETAs or $c_j$? Comment in light of part 4.

(f) Removing observations found in part 4, refit the model. Do the coefficients/significance change? Write a coherent summary of your findings.

7. Agresti 6.4. The marijuana.txt dataset can be downloaded from the course webpage.

8. Dixon and Massey (1983) present data on 200 men taken from the Los Angeles Heart Study. The data are in heart.sas; ignore the last column named as garbage. There are 7 variables from left to right: age (AG), systolic blood pressure (S), diastolic blood pressure (D), cholesterol (Ch), height (H), weight (W), and whether a coronary incident occurred (CNT) (1=incident occurred in previous decade, 0=not). There were 26 incidents among the men. According to our rule of thumb we should have $26/10 = 2.6 \approx 2-3$ predictors at most in the final model.

(a) Use backwards elimination and stepwise procedures to find final models using defaults SLENTRY=SLSTAY=0.05. Does your final adhere to the rule of thumb?

(b) For your final model, prepare plots of $r_i$ vs. each predictor with loess smooths superimposed, and $c_i$ vs. $i$ and comment on model fit and influential diagnostics.

(c) Interpret your final model

(d) Discuss the final model’s predictive ability using options available in proc logistic.