

Homework #4

Please email your completed assignment to biostat208hw@gmail.com by 10:30 AM on Tuesday, March 8, 2016.

The first two exercises focus on interpretation of single-predictor logistic regression models fitted to example data from the HERS study. Recall the HERS was a randomized trial of Estrogen plus Progestin for prevention of coronary heart disease (CHD) events in postmenopausal women (Hulley S, *et al*; *JAMA*, 1998). Here we consider baseline data from 1055 women without diabetes assigned to the placebo arm, and focus on the relationship between a measure of hypertension (`hypt` a binary indicator of a baseline systolic blood pressure (SBP) level of greater than 140 mmHg), exercise (`exer3`, defined as a binary indicator of self-reported exercise of at least 3 times per week) and body mass index (BMI, kg/m²).

1. The output below displays a fitted logistic regression model for the relationship between hypertension (`hypt`) and exercise (`exer3`), both defined as described above. As discussed in lecture, the `coef` option to the `logistic` command requests that the results be reported in terms of the regression coefficients rather than the estimated odds ratio for `exer3`. Answer the following questions about the output:
 - A. Using the estimated coefficients, write down the logistic equation for the model in both nonlinear (probability) and linear (log odds) form.
 - B. Give interpretations for each of the coefficients (`_cons` and `exer3`).
 - C. Exponentiate the coefficient for `exer3` and the lower and upper bounds for its 95% confidence interval.
 - D. State the interpretation of the estimated odds ratio for `exer3`.
 - E. Interpret the z statistic and associated P -value for the coefficient of `exer3`, and contrast these results with the 95% confidence interval in the output.

```
. logistic hypt exer3, coef
```

```
Logistic regression               Number of obs   =       1055
                                LR chi2(1)         =        10.12
                                Prob > chi2         =        0.0015
Log likelihood = -667.38065       Pseudo R2        =        0.0075
```

hypt	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
exer3	-.4267843	.1352317	-3.16	0.002	-.6918335	-.1617351
_cons	-.5185876	.0831043	-6.24	0.000	-.681469	-.3557063

2. The output below displays a fitted logistic regression model for the relationship between hypertension (`hypt`) and BMI (`bmi`), both defined as described above. Again, the output is presented using the `coef` option to the `logistic` command.
- Using the estimated coefficients, write down the logistic equation for the model in both nonlinear (probability) and linear (log odds) form.
 - Give interpretations for each of the coefficients (`_cons` and `bmi`).
 - Compute the corresponding odds ratio for the coefficient of `bmi`, and the associated 95% confidence interval.
 - State the interpretation of the estimated odds ratio for `bmi`.
 - Use the output to compute the odds ratio for a 30-unit increase in `bmi`, and the associated 95% confidence interval.
 - Use the linear form of the logistic model in part A to compute the log odds of hypertension for a hypothetical woman with a BMI of 50 kg/m². Repeat for a woman with a BMI of 20 kg/m². Use the two log odds obtained to calculate the odds ratio for an increase in BMI of 30 kg/m².
 - Repeat part F using the nonlinear (probability) form of the logistic equation for the two different BMI values, and (rather than the odds ratio) estimate the “relative risk” comparing the outcome probability for a BMI of 50 kg/m² with that for a BMI of 20 kg/m². How could we obtain the 95% confidence interval for this estimate?

```
. logistic hypt bmi, coef
```

```
Logistic regression               Number of obs   =       1055
                                LR chi2(1)         =         9.35
                                Prob > chi2         =         0.0022
Log likelihood = -667.76922       Pseudo R2        =         0.0070
```

hypt	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
bmi	.0376171	.012309	3.06	0.002	.0134918	.0617423
_cons	-1.742098	.3529975	-4.94	0.000	-2.43396	-1.050235

For the following question use the data set `icu.dta`, available on the course site. These data were collected on 200 patients who were part of a larger study investigating in-hospital survival of patients admitted to an ICU. A description of the study and variables is included on the next page. The analysis in the exercise can be considered part of a larger effort to develop a predictive risk score model for mortality following admission to the ICU.

3. Use the `logistic` command to fit a regression model relating the binary outcome variable `sta` (vital status) to the variables described below:
 - `crn` (a binary indicator of history of chronic renal failure),
 - `typ` (a binary indicator of type of admission),
 - `can` (a binary indicator of presence of cancer being related to admission),
 - `loc` (a 3-level categorical variable describing level of consciousness),
 - `agecat` (a 4-category representation of age),
 - `syscat` (a 4-category representation of systolic blood pressure).

Note that although continuous measures of age and systolic blood pressure are available, categorical versions were chosen to increase interpretability and to deal with possible nonlinearities in the relationships of these variables with the log odds of the outcome. Use the following statements to create the categorical representations of age and systolic blood pressure. Note that the cut-points chosen for both variables roughly correspond to the quartiles in the sample.

```
recode age (0/47=1) (48/63=2) (64/72=3) (73/max=4), gen(agecat)
recode sys (0/110=1) (111/130=2) (131/150=3) (150/max=4), gen(syscat)
```

Answer the following questions about the output produced:

- A. Give an interpretation of the reported odds ratios for `crn` and `typ`.
- B. Comment on the results for the variable `loc`, focusing on the following questions: Why was no odds ratio estimated for the 2nd level of this variable? Can you think of a way to address this problem?
- C. Conduct a test for linear trend of the log odds of the outcome with increasing age, using the procedures described in section 4.3.5 of the book, and illustrated in lab #3.
- D. Estimate the odds ratio comparing the odds of the outcome between the 3rd and 2nd quartiles of age, and report this along with the corresponding 95% confidence interval.
- E. Conduct 2 likelihood ratio tests to investigate the separate contributions of age (as represented by `agecat`) and systolic blood pressure (as represented by `syscat`) to the model including all the predictors described above, and report the results.

icu.dta: 200 observations, 21 variables

Description:

The ICU data set consists of a sample of 200 subjects who were part of a much larger study on survival of patients following admission to an adult intensive care unit (ICU). The major goal of this study was to develop a logistic regression model to predict the probability of survival to hospital discharge of these patients and to study predictors associated with ICU mortality.

Description	Codes/Values	Variable name
Identification Code	ID number	id
Vital Status	0 = Lived, 1 = Died	sta
Age	Age in years	age
Gender	0 = Male, 1 = Female	sex
Ethnicity	1 = White 2 = Black 3 = Other	race
Service at admission	0 = Medical 1 = Surgical	ser
Cancer part of problem	0 = No, 1 = Yes	can
History of chronic renal failure	0 = No 1 = Yes	crn
Infection at admission	0 = No 1 = Yes	inf
CPR prior to admission	0 = No, 1 = Yes	cpr
Syst. blood pressure	mm Hg	sys
Heart rate	beats/min	hra
Previous admission to ICU within 6 months	0 = No 1 = Yes	pre
Type of admission	0 = Elective 1 = Emergency	typ
Major bone fracture at admission	0 = No 1 = Yes	fra
PO2 from initial blood gases	0 $\geq$ 60 1 < 60	po2
PH from initial blood gases	0 > 7.25 1 $\leq$ 7.25	ph
PCO2 from initial blood gases	0 > 45 1 $\leq$ 45	pco
Bicarbonate from initial blood gases	0 > 18 1 $\leq$ 18	bic
Creatinine from initial blood gases	0 < 2.0 1 $\geq$ 2.0	cre
Level of consciousness	0 = No coma/stupor 1 = Deep stupor 2 = Coma	loc

REFERENCES

Hosmer and Lemeshow, *Applied Logistic Regression*, Wiley, (1989)
Lemeshow, S., Teres, D., Avrunin, J. S., Pastides, H. (1988). Predicting the Outcome of Intensive Care Unit Patients. *Journal of the American Statistical Association*, **83**, 348-356.