

Midterm Exam III

Instructions: This is a close-book and close-notes exam. You are only allowed to bring three cheat sheets. Arrange your time wisely and show all your work. Details and explanations are required in order to receive partial/full credit. Note that the total score of this exam is 110, including ten extra credits

1. We consider the Stanford heart transplant data. The binary (1 for 'dead' and 0 for 'alive') response death (Y) indicate the survivorship of patients on the waiting list for the Stanford heart transplant program. Four predictors are included. a brief description is given below.

```

-----
age (x1):          age-48 years
year (x2):         year of acceptance (in years after 1 Nov 1967);
surgery (x3):     prior bypass surgery 1=yes 0=no
transplant (x4):  received transplant 1=yes 0=no
-----

```

A number of logistic regression models are fit in order to study the effect of `transplant` (X_4) on the survivorship.

$$\text{Model A: } \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_4.$$

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-0.889	0.217	-4.10	4.1e-05	***
transplant1	1.518	0.333	4.56	5.2e-06	***

Null deviance: 235.62 on 171 degrees of freedom
Residual deviance: 213.44 on 170 degrees of freedom

Unscaled Covariance Matrix for Betas:

	(Intercept)	transplant1
(Intercept)	0.04703	-0.04703
transplant1	-0.04703	0.11092

$$\text{Model B: } \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4.$$

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-0.3381	0.9579	-0.35	0.72408	
age	0.0145	0.0185	0.78	0.43349	
year	-0.3661	0.1036	-3.53	0.00041	***
surgery	-0.6075	0.4867	-1.25	0.21200	
transplant1	1.8221	0.3742	4.87	1.1e-06	***

Null deviance: 235.62 on 171 degrees of freedom
Residual deviance: 194.42 on 167 degrees of freedom

$$\text{Model C: } \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_2 x_2 + \beta_4 x_4.$$

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.326	0.367	0.89	0.37381
year	-0.394	0.102	-3.88	0.00010 ***
transplant1	1.797	0.369	4.87	1.1e-06 ***

Null deviance: 235.62 on 171 degrees of freedom
 Residual deviance: 196.58 on 169 degrees of freedom

Unscaled Covariance Matrix for Betas:

	(Intercept)	year	transplant1
(Intercept)	0.13462	-0.02915	-0.01841
year	-0.02915	0.01031	-0.01195
transplant1	-0.01841	-0.01195	0.13632

$$\text{Model D: } \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_1 + \beta_4 x_4.$$

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.0356	0.8776	-2.32	0.020 *
age	0.0251	0.0184	1.36	0.173
transplant1	1.5158	0.3352	4.52	6.1e-06 ***

Null deviance: 235.62 on 171 degrees of freedom
 Residual deviance: 211.51 on 169 degrees of freedom
 AIC: 217.5

Unscaled Covariance Matrix for Betas:

	(Intercept)	age	transplant1
(Intercept)	0.77019	-0.0156420	-0.0580606
age	-0.01564	0.0003386	0.0002277
transplant1	-0.05806	0.0002277	0.1123259

$$\text{Model E: } \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_1 + \beta_4 x_4 + \beta_5 \cdot x_1 \cdot x_4.$$

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.2707	0.9947	-0.27	0.786
age	-0.0138	0.0217	-0.63	0.526
transplant1	-3.3921	1.8402	-1.84	0.065 .
age:transplant1	0.1084	0.0401	2.70	0.007 **

Null deviance: 235.62 on 171 degrees of freedom
 Residual deviance: 203.50 on 168 degrees of freedom

Unscaled Covariance Matrix for Betas:

	(Intercept)	age	transplant1	age:transplant1
(Intercept)	0.9894	-0.0210994	-0.98940	0.0210994
age	-0.0211	0.0004725	0.02110	-0.0004725
transplant1	-0.9894	0.0210994	3.38624	-0.0725411
age:transplant1	0.0211	-0.0004725	-0.07254	0.0016118

PART I:

- (a) (10 points) Concerning Model A, give a 95% confidence interval (CI) for the odds ratio of death between patients who had heart transplant and those who did not.
- (b) (10 points) Concerning Model B, give a 95% confidence interval (CI) for the odds ratio of death between patients who had heart transplant and those who did not, after adjusting for other covariates.
- (c) (3 points) Based on the comparison of part (a) and part (b), comment on existence of any confounding effect on **transplant** (X_4).
- (d) (12 points) Based on Model A, give a 95% confidence interval (CI) to predict the death rate, i.e., $P(y = 1)$ of patients who did NOT have heart transplant (**Hint:** namely, $X_4 = 0$).
- (e) (10 points) Concerning Model B, apply the likelihood ratio test (LRT) to test $H_0 : \beta_1 = \beta_3 = 0$.

PART II: Suppose, in particular, it is of interest to test whether the effect of **transplant** (X_4) would vary with **age** (X_1). That is, their interaction is under concern.

- (a) (10 points) Concerning Model D, construct a 95% confidence interval for the the odds ratio of death between patients who had heart transplant and those who did not, holding other covariates fixed at ($X_1 = 40, X_2 = 2, X_3 = 0$).
- (b) (10 points) Concerning Model E, construct a 95% confidence interval for the the odds ratio of death between patients who had heart transplant and those who did not, holding other covariates fixed at ($X_1 = 40, X_2 = 2, X_3 = 0$).
- (c) (5 points) Using Model E, perform a statistical test to see whether interaction between **transplant** and **age** really exists. (Hint: either LRT or Wald test is fine.)

2. The model fit given below is based on a data set that comes from the Veterans' Administration Lung Cancer study. (Reference: D Kalbfleisch and RL Prentice (1980), *The Statistical Analysis of Failure Time Data*. Wiley, New York.) Suppose that we would like to compare the survival rates among lung cancer patients with different cell types. The binary response Y is coded as 1 for death and 0 for survivor. And the categorical variable **celltype** (Z) has four levels: **squamous**, **smallcell**, **adeno**, and **large**. Thus three dummy variables are defined, using the reference cell scheme.

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	2.048	0.531	3.85	0.00012 ***
smallcell	0.660	0.799	0.83	0.40832
adeno	1.210	1.149	1.05	0.29223
large	1.210	1.149	1.05	0.29223

Null deviance: 66.405 on 136 degrees of freedom
 Residual deviance: 64.429 on 133 degrees of freedom

	(Intercept)	smallcell	adeno	large
(Intercept)	0.2823	-0.2823	-0.2823	-0.2823
smallcell	-0.2823	0.6378	0.2823	0.2823
adeno	-0.2823	0.2823	1.3207	0.2823
large	-0.2823	0.2823	0.2823	1.3207

Based on the above fit, answer the following questions.

- (a) (6 points) Give the definitions of three dummy variables (denote them by, say, Z_1 , Z_2 and Z_3) that are associated with the above model. Then write down the logistic model. Which cell type is used as the *baseline*?
 - (b) (10 points) Construct a 95% CI for the odds ratio of death that compares the **large** group and the **squamous** group.
 - (c) (14 points) Construct a 95% CI for the odds ratio of death that compares the **large** group and the **smallcell** group.
3. (10 points) Consider the grouped data $\{(m_k, y_k, \mathbf{x}_k) : k = 1, \dots, K\}$, where m_k is the total number of observations having the same k th covariate pattern \mathbf{x}_k and y_k is the number of them who have response value 1. Working, however, with the null model that includes the intercept only:

$$\log\left(\frac{\pi_k}{1 - \pi_k}\right) = \beta_0, \quad (1)$$

Write down the log-likelihood function. Then show that the maximum likelihood estimate (MLE) of

$$\hat{\beta}_0 = \log\left(\frac{p}{1 - p}\right), \text{ with } p = \frac{\sum_{k=1}^K y_k}{\sum_{k=1}^K m_k} \text{ being the sample proportion of 1's.}$$

(Hint: Here is one easy way to approach. Note that, according to Model (1), π_k must be a constant that does not depend on the subscript k . Thus let $\pi_k = \pi$ in the log-likelihood function, proceed to find the MLE of π first, and then transform it back to obtain $\hat{\beta}_0$.)

Solution

1. PART I

- (a) The question asks for 95% CI for $OR = \exp(\beta_1)$. First get CI for β_1 : $1.518 \pm 1.96 \times 0.333 \implies (0.865, 2.171)$ and hence CI for $\exp(\beta_1)$ is $(2.376, 8.764)$.
- (b) It asks for CI for $OR = \exp(\beta_4)$: $(2.970, 12.878)$.
- (c) The CI's in part (a) and (b) are not quite different, thus perhaps the confounding effect is not a big concern, assuming no interaction between X_4 with other covariates.
- (d) Plug in $X_4 = 0$ into Model A, it can be found that $\pi_i = \exp(\beta_0)/(1 + \exp(\beta_0))$. First find CI for β_0 : $(-1.314, -0.464)$. Then CI for π_i is

$$\left(\frac{\exp(-1.314)}{1 + \exp(-1.314)}, \frac{\exp(-0.464)}{1 + \exp(-0.464)} \right) \implies (0.212, 0.386).$$

- (e) With Model B being the *whole* or *full* model and Model C being the reduced model under H_0 , $LRT = 2.16$, which is less than $\chi_{0.95}^2(2) = 5.991$. Thus we cannot reject the null H_0 .

PART II

- (a) Since Model D is an additive model, the wanted $OR = \exp(\beta_4)$. A Wald 95% CI can be found as $(2.360, 8.783)$.
- (b) Model E is an interaction model, the wanted $OR = \exp(\beta_4 + 40\beta_5)$. First find 95% CI for $(\beta_4 + 40\beta_5)$:

$$\begin{aligned} \hat{\beta}_4 + 40\hat{\beta}_5 &= -0.0138 + 40 \times 0.1084 = 0.944 \\ \text{var}(\hat{\beta}_4 + 40\hat{\beta}_5) &= \text{var}(\hat{\beta}_4) + 40^2 \cdot \text{var}(\hat{\beta}_5) + 2 \times 40 \times \text{cov}(\hat{\beta}_4, \hat{\beta}_5) \\ &= 3.386 + 1600 \times 0.0016118 + 80 \times (-0.07254) \\ &= 0.1619 \end{aligned}$$

Hence we have $0.944 \pm 1.96 \times \sqrt{0.1619} \implies (0.1554, 1.7324)$.

Therefore, CI for OR: $(\exp(0.1554), \exp(1.7324)) \implies (1.168, 5.654)$.

- (c) Using the Wald test given in the table, $z = 2.70$ with p-value 0.007. Thus we cannot reject the null. Interaction is significant.
2. (a) Define $Z_1 = 1$ if a patient belongs to the **smallcell** group and 0 otherwise; $Z_2 = 1$ if a patient belongs to the **adeno** group and 0 otherwise; $Z_3 = 1$ if a patient belongs to the **large** group and 0 otherwise. In this way, the **squamous** group would be left as the *baseline*. The corresponding logistic model is $\text{logit}(\pi_i) = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \beta_3 z_{i3}$.
- (b) The OR would be $\exp(\beta_3)$ in this case. The 95% CI is $(0.353, 31.88)$.
- (c) The OR would be $\exp(\beta_3 - \beta_1)$ with CI $(0.171, 17.53)$.

3. Working with the grouped data, we have $y_k \sim \text{Binomial}(m_k, \pi_k)$ and hence the log-likelihood function would be

$$l = \log L = \sum_{k=1}^K \log \binom{m_k}{y_k} + \sum_{i=1}^n \{y_k \log(\pi_k) + (m_k - y_k) \log(1 - \pi_k)\}. \quad (2)$$

Now plug in $\pi_k = \exp(\beta_0)/(1 + \exp(\beta_0)) = \pi$ and maximize it with respect to π :

$$0 \equiv \frac{\partial l}{\partial \pi} = \sum_{k=1}^K \frac{y_k - m_k \pi}{\pi(1 - \pi)}.$$

Thus $\hat{\pi} = \sum y_k / \sum m_k = p$ and hence $\hat{\beta}_0 = \log(p/(1 - p))$.