Department of Mathematics

The University of Toledo

Master of Science Degree
Comprehensive Examination

Applied Statistics

April 22, 2017

Instructions

Do all four problems.

Show all of your computations.
Prove all of your assertions or quote appropriate theorems.
Books, notes, and calculators may be used.
This is a three hour test.

MS Applied Exam

Categorical Data Analysis

Below is the data on Admission to Berkeley.

	Whether Admitted			
	Male		Female	
Department	Yes	No	Yes	No
1	512	313	89	19
2	353	207	17	8
3	120	205	202	391
4	138	279	131	244
5	53	138	94	299
6	22	351	24	317
Total	1198	1493	557	1278

Figure 1:

1. (3 pts) Calculate and **interpret** the marginal odds ratio between admission and gender. Please show the 95% CI.
2. (3 pts) Calculate and **interpret** the odds ratio between admission and gender for Department 1 and 2, respectively. Please give their 95% CI as well.
3. Below are some R code and their output. Answer the questions according to the output:

```
dat <- cbind(yes=c(512,353,120,138,53,22,89,17,202,131,94,24),
             no=c(313,207,205,279,138,351,19,8,391,244,299,317))
dept = factor(rep(1:6,2))
gender = rep(c("M","F"),each=6)

fit1 <- glm(dat~gender,family=binomial(logit))
summary(fit1)$coef
```

```
##               Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -0.8304864 0.05077208 -16.35715 3.868379e-60
## genderM      0.6103524 0.06389305   9.55272 1.263352e-21
```

```
exp(fit1$coef)
```

```
## (Intercept)     genderM
##   0.4358372   1.8410800
```

```r
fit2 <- glm(dat~gender*dept,family=binomial(logit))
fit2$coef
```

```
##   (Intercept)       genderM         dept2         dept3         dept4 
##     1.5441974    -1.0520760    -0.7904256    -2.2046373    -2.1661683 
##         dept5         dept6 genderM:dept2 genderM:dept3 genderM:dept4 
##    -2.7013462    -4.1250453     0.8320534     1.1769976     0.9700888 
## genderM:dept5 genderM:dept6 
##     1.2522630     0.8631801
```

```r
exp(fit2$coef)
```

```
##   (Intercept)       genderM         dept2         dept3         dept4 
##    4.68421053    0.34921205    0.45365169    0.11029053    0.11461595 
##         dept5         dept6 genderM:dept2 genderM:dept3 genderM:dept4 
##    0.06711510    0.01616276    2.29803272    3.24461787    2.63817862 
## genderM:dept5 genderM:dept6 
##    3.49825046    2.37068781
```

(1) (2 pts) What is the reference level for `gender` in `fit1` and `fit2`?

(2) (2 pts) What is the reference level for `dept` in `fit2`?

(3) (4 pts) In `fit1`, how to interpret `genderM=0.610`? Can you find the 95% CI for it?

(4) (3 pts) In `fit1`, how to interpret the intercept, i.e., `intercept=-0.830`?

(5) (5 pts) Can you write down the model for `fit1` with appropriate coefficient and variable names?

(6) (7 pts) Use the output of `fit2` to answer the question: find the log odds for a male applicant applying **Department 1** and log odds for a female applicant applying **Department 1** and find the log odds ratio and the odds ratio.

(7) (7 pts) Use the output of `fit2` to answer the question: find the log odds for a male applicant applying **Department 2** and log odds for a female applicant applying **Department 2** and find the log odds ratio.

(8) (5 pts) Can you find the **odds ratio** between admission and gender for **Department 3** using `fit2`? Show your work and verify it using the numbers in the Figure 1.

Linear Regression

You are given a dataset with one response variable y , and two predictor variables x_1, x_2, with $n = 5$ observations. You are to fit the following multiple linear regression model, *without an intercept*:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, i = 1, 2, ..., 5$$

(1) (10 pts) Write out the matrix form of the multiple linear regression, including the error assumptions as well. Indicate the dimensions for all matrices.

(2) (5 pts) To estimate the coefficients, we minimize the sum of squares $\sum_{i=1}^{n}(y_i - \beta_1 x_{i1} - \beta_2 x_{i2})^2$ and derive the following estimating equations. Please show your work to verify the below estimating euqations.

$$
\begin{aligned}
\sum_i x_{i1} y_i &= \beta_1 \sum_i x_{i1}^2 + \beta_2 \sum_i x_{i1} x_{i2} \\
\sum_i x_{i2} y_i &= \beta_1 \sum_i x_{i1} x_{i2} + \beta_2 \sum_i x_{i2}^2
\end{aligned}
$$

(3) (5 pts) If we rewrite the minimization problem using matrices, the solution is as below

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Show that $\hat{\boldsymbol{\beta}}$ is an unbiased estimate of $\boldsymbol{\beta}$ and derive the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$

(4) (5 pts) Write out the hat matrix $\mathbf{H}$ in terms of $\mathbf{X}$. What is the dimension of the hat matrix $\mathbf{H}$ for this model?

Multivariate Statistics

Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{20}$ be a random sample of size $n = 20$ from an $N_6(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ population.

1. Specify the distribution of the following:

(a) (3 pts) $(\mathbf{X}_1 - \boldsymbol{\mu})^T(\boldsymbol{\Sigma})^{-1}(\mathbf{X}_1 - \boldsymbol{\mu})$
(b) (3 pts) $\bar{\mathbf{X}}$
(c) (3 pts) $\sqrt{n}(\bar{\mathbf{X}} - \boldsymbol{\mu})$
(d) (3 pts) $V_1 = \mathbf{X}_1 - \mathbf{X}_2 + \mathbf{X}_3 - \mathbf{X}_4$
(e) (3 pts) $V_2 = \mathbf{X}_1 + \mathbf{X}_2 + \mathbf{X}_3 + \mathbf{X}_4$
(f) (5 pts) Joint distribution of V_1 and V_2 defined above

Nonparametric Statistics

(20 points) In an early study of the effects of a strong magnetic field on the development of mice, 7 cages, each containing 3 albino female mice were subjected for a period of ten days to a magnetic field. 21 other mice housed in 7 similar cages were not placed in a magnetic field and served as controls. The following table shows the weight gains, in grams, for each of the cages.

Magnetic Field Present: 22.8 10.2 20.8 27.0 19.2 10.4 14.2

Magnetic Field Absent: 23.6 31.0 19.5 26.2 26.5 25.2 24.5

(a) (10 pts) State a nonparametric model, define hypotheses, and carry out a test at level $\alpha = 0.10$ which will enable you to decide whether there is a significant difference in weight gain between these two groups. Find the exact p-value and use it to make your decision.

(b) (10 pts) Repeat part a using the normal approximation. Do you get the same conclusion from (a)?