**BIOS6643 Homework 2 Due Friday, September 23, 2016**

Practice questions (not to turn in):

1. Consider using PROC GLM in SAS to fit a regression model. If one of the predictors is gender, coded as an indicator variable (e.g., ‘1’ for Female and ‘0’ for Male), you will essentially get the same model fit whether or not you put this variable into the CLASS statement. Thus, although it is clearly not a continuous variable, we can treat it as such when fitting the model. Briefly describe why this is the case.

Because the indicator variable will only have a 0 if male and 1 if female thus can be counted as both continuous and categorical and regardless will be fitting the model the same way. The data also will only have 0s and 1s and the outcome will be regressed on the 0s and 1s only thus whether specifying it as categorical or continuous will not matter.

1. Review Section 3.6.3 in the GLM I slide set and course notes.
   1. Write full-rank and less-than-full-rank models if there is a group variable with 4 levels (i.e., 4 groups) and time is still treated as a continuous variable. How many columns are in X for each approach?

Full Rank Model

Less Than Full Rank Model

There would be 10 columns

* 1. If time points are unequally spaced then would it be appropriate to treat time as a class variable? Explain.

No, because it would cause the model to have too many columns created thus the interpretation of each beta parameters in comparison to the reference group would not be ideal.

1. Complete the practice question on the bottom of page 29 in the GLM II slides.

=LH

1. On slide 24 of the GLM II slide set, the distribution of was shown, which was derived using the linear form result (see slide 21, GLM I slides). Derive the distribution of  if  form some vector **a**. Note that the result is given on slide 25 and that this can also be completed using the linear form result. The proof is short.

=

1. We’ve discussed how in some cases, the approach of simplifying a general linear model that has class variables up front will yield the same result as using a g-inverse. In particular, setting the highest level(s) of factor(s) to 0 essentially yields the same results as they way SAS computes the g-inverse. In the set-to-0 approach, we know that estimates for levels other than the level that did not have an indicator reflect comparisons to that level (i.e., the one without the indicator is the ‘reference level’). Using estimability and the less-than-full-rank model, show how we know this to be true. [Hint: consider estimability of κi–κj in the one-way effects model.]

Because using the generalized inverse method of dropping the linearly dependent columns from left to right, L =LH form of equation will be satisfied for κi–κj, in general.

To turn in:

1. Top race times by age were recorded for the Bolder Boulder 10K race (top 5 times for each individual year age in the 1995 race). We are going to model race time as a function of age and gender, including linear and quadratic terms for age, as well as age×gender and age2×gender interaction terms, for 30 to 60 year-old males and females. Here, we don’t have a random sample; these are extreme values since they are the fastest times for each age, but we are more concerned with curve fitting than inference.
   1. Write the model in terms of a single outcome (i.e., not in matrix form).
   2. Write **β** and the first 10 rows of **X** for the matrix model (get the data from the course web site).

X =

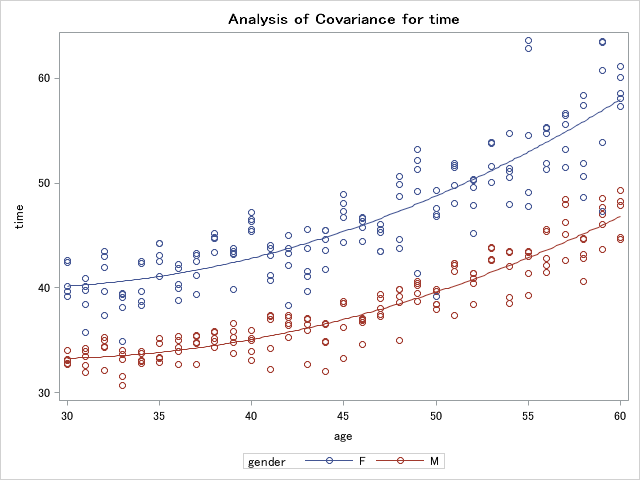
| **Design Points** | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Observation Number** | **time** | **Column Number** | | | | | | | | |
|  |  |  |  |  |  |  |  |  |
| **1** | 32.7 | 1 | 30 | 900 | 0 | 1 | 0 | 30 | 0 | 900 |
| **2** | 32.8 | 1 | 30 | 900 | 0 | 1 | 0 | 30 | 0 | 900 |
| **3** | 33.1 | 1 | 30 | 900 | 0 | 1 | 0 | 30 | 0 | 900 |
| **4** | 33.2 | 1 | 30 | 900 | 0 | 1 | 0 | 30 | 0 | 900 |
| **5** | 34.0 | 1 | 30 | 900 | 0 | 1 | 0 | 30 | 0 | 900 |
| **6** | 39.2 | 1 | 30 | 900 | 1 | 0 | 30 | 0 | 900 | 0 |
| **7** | 39.6 | 1 | 30 | 900 | 1 | 0 | 30 | 0 | 900 | 0 |
| **8** | 40.2 | 1 | 30 | 900 | 1 | 0 | 30 | 0 | 900 | 0 |
| **9** | 42.4 | 1 | 30 | 900 | 1 | 0 | 30 | 0 | 900 | 0 |
| **10** | 42.6 | 1 | 30 | 900 | 1 | 0 | 30 | 0 | 900 | 0 |

* 1. Analyze the data with PROC GLM. Write the fitted functions for 10K race times by age for the top males and females.

**Model**

**Fitted-Function**

* 1. Graph the data using either SAS or R. Use different symbols for males and females and superimpose the fitted functions.



1. Show that  satisfies the normal equations. (Here, tilde is used to indicate that the beta estimate may not be unique.)

Normal Equation

Substitute

Hence

Thus:

1. Consider a study or experiment that has two factors (e.g., group and time); each factor has 4 levels and will be treated as a class variable. We will create a model for response ‘y’ as a function of *group*, *time* and *group\*time*; there are 3 replicates for each group-time combination.
   1. How many columns are in **X** for the less-than-full-rank model?

, group\*time k=1…16

There would be 25 columns

* 1. If you were to write a full-rank statistical model for these data, how many parameters would there be? (I.e., how many columns are in **X** for a full-rank model?)

The full rank model would only have 16 columns.

1. For the Myostatin data, note that the population mean for the myostatin group at 48 hours is  for the one-way model (see notes). Write the population means for the following.
   1. Myostatin group at 48 hours; means model.

5.25125.

* 1. Myostatin group at 48 hours; two-way effects model.

, 5.25125.

* 1. Myostatin group, difference between 48 and 72 hours, one-way effects model.

, 0-1.473 = -1.473

* 1. Myostatin group, difference between 48 and 72 hours, two-way effects model.

, -1.473.

1. Consider a 3×2 factorial experiment with 2 replicates in each treatment combination. The data will be analyzed using the model , *i*=1,2,3; *j*=1,2; *k*=1,2 (*k* denotes the replicate).
   1. Write the **β** and **X** matrices for the general linear model for the *effects model* shown above.

**X =**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 1 | 1 | 0 | 0 | 1 | 0 |
| 1 | 1 | 0 | 0 | 1 | 0 |
| 1 | 1 | 0 | 0 | 0 | 1 |
| 1 | 1 | 0 | 0 | 0 | 1 |
| 1 | 0 | 1 | 0 | 1 | 0 |
| 1 | 0 | 1 | 0 | 1 | 0 |
| 1 | 0 | 1 | 0 | 0 | 1 |
| 1 | 0 | 1 | 0 | 0 | 1 |
| 1 | 0 | 0 | 1 | 1 | 0 |
| 1 | 0 | 0 | 1 | 1 | 0 |
| 1 | 0 | 0 | 1 | 0 | 1 |
| 1 | 0 | 0 | 1 | 0 | 1 |

* 1. What is *r*(**X**)?

**4**

* 1. Are the following estimable? Justify your response.
     1. 

H matrix

|  | | | | | |
| --- | --- | --- | --- | --- | --- |
| 0.5454545 | 0.1818182 | 0.1818182 | 0.1818182 | 0.2727273 | 0.2727273 |
| 0.1818182 | 0.7272727 | -0.272727 | -0.272727 | 0.0909091 | 0.0909091 |
| 0.1818182 | -0.272727 | 0.7272727 | -0.272727 | 0.0909091 | 0.0909091 |
| 0.1818182 | -0.272727 | -0.272727 | 0.7272727 | 0.0909091 | 0.0909091 |
| 0.2727273 | 0.0909091 | 0.0909091 | 0.0909091 | 0.6363636 | -0.363636 |
| 0.2727273 | 0.0909091 | 0.0909091 | 0.0909091 | -0.363636 | 0.6363636 |

LH = (1 0 0 0 0 0)H = (0.5454545 0 0 0 0 0) != (1 0 0 0 0 0)

Not estimable.

* + 1. 

LH = (0 0 0 1 0 0)H = (0.1818182 -0.272727 -0.272727 0.72727272 0.90909091 0.90909091) != L = (0 0 0 1 0 0 0)

Not estimable.

* + 1. 

LH = (1 0 0 1 0 0)H =

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| (0.7272727 | -0.090909 | -0.090909 | 0.9090909 | 0.3636364 | 0.3636364) |

!= L = (1 0 0 1 0 0 0)

Not Estimable.

* + 1. 

LH = (1 0 0 1 0 1)H =

(0.5454545+0.1818182+0.2727273 0.1818182-0.2727272+0.0909091 --- 0.2727273+0.0909091+0.63636363)

= (1 0 0 1 0 1) = L = (1 0 0 1 0 1 0)

Estimable.

* 1. For the bread data (Neter, p. 686; posted on web page) determine the following using SAS PROC IML, R, or other software, based on the less-than-full-rank model written above.
     1. 

| **betahatx** |
| --- |
| 27.818182 |
| 2.2727273 |
| 25.272727 |
| 0.2727273 |
| 12.909091 |
| 14.909091 |

* + 1. ii. 

| **pred\_valuesx** |
| --- |
| 43 |
| 43 |
| 45 |
| 45 |
| 66 |
| 66 |
| 68 |
| 68 |
| 41 |
| 41 |
| 43 |
| 43 |

* + 1. iii. S.E. ()

1.8929694

Using PROC GLM to check and verify

| **Parameter** | **Estimate** |  | **Standard Error** | **t Value** | **Pr > |t|** |
| --- | --- | --- | --- | --- | --- |
| **Intercept** | 43.00000000 | B | 1.89296945 | 22.72 | <.0001 |
| **a 1** | 2.00000000 | B | 2.31840462 | 0.86 | 0.4134 |
| **a 2** | 25.00000000 | B | 2.31840462 | 10.78 | <.0001 |
| **a 3** | 0.00000000 | B | . | . | . |
| **b 1** | -2.00000000 | B | 1.89296945 | -1.06 | 0.3216 |
| **b 2** | 0.00000000 | B | . | . | . |

SAS CODES FOR Q1

libname temp "C:\Users\kimchon\Downloads";

**data** temps;

set temp.bb\_data;

agesq = age\*age;

**run**;

\*GLM model also spits out the graphic;

**proc** **glm** data = temps;

class gender;

model time = age age\*age gender age\*gender age\*age\*gender /solution xpx i ss3;

**run**; **quit**;

\*save the matrix for 1B;

ods output designpoints=xmatrix;

**proc** **glmmod** data=temps;

class gender;

model time = age agesq gender age\*gender agesq\*gender ;

**run**;

SAS CODES FOR Q5

**proc** **iml**;

\*LESS THAN FULL RANK MODEL;

x={

**1** **1** **0** **0** **1** **0**, **1** **1** **0** **0** **1** **0**,

**1** **1** **0** **0** **0** **1**, **1** **1** **0** **0** **0** **1**,

**1** **0** **1** **0** **1** **0**, **1** **0** **1** **0** **1** **0**,

**1** **0** **1** **0** **0** **1**, **1** **0** **1** **0** **0** **1**,

**1** **0** **0** **1** **1** **0**, **1** **0** **0** **1** **1** **0**,

**1** **0** **0** **1** **0** **1**, **1** **0** **0** **1** **0** **1**

} ;

y={

**47**, **43**, **46**, **40**, **62**, **68**, **67**, **71**, **41**, **39**, **42**, **46**

};

l={

**1** **0** **0** **1** **0** **1**

};

xt=t(x);

xtx=xt\*x;

xtxginv=ginv(xtx);

px=x\*xtxginv\*xt;

h=xtxginv\*xt\*x;

lh=l\*h;

pred\_valuesx=px\*y;

betahatx=xtxginv\*xt\*y;

print pred\_valuesx;

print betahatx;

print lh;

print h;

\*determine rank of x;

rankGInv = round(trace(ginv(x)\*x));

print rankGInv;

e = echelon(x);

print e;

\*echelon form has 4 rows with 0 zeros thus rank=4;

\*SE(u+a3+tau2) estimation;

yt=t(y);

ident=I(**12**);

mse=(yt\*(ident-px)\*y)/(**12**-**4**);

lt=t(l);

lxtxlt = l\*xtxginv\*lt;

se = sqrt(mse\*lxtxlt);

print se;

\*SE(u+a3+tau2) estimation using GLM to make sure answers match up;

**data** bread;

input score a b rep;

label a ='Height'

b ='Width'

rep ='Exp. unit';

cards;

47 1 1 1

43 1 1 2

46 1 2 1

40 1 2 2

62 2 1 1

68 2 1 2

67 2 2 1

71 2 2 2

41 3 1 1

39 3 1 2

42 3 2 1

46 3 2 2

;

**run**;

**proc** **glm** data=bread; class a b; model score = a b / E solution xpx; **run**;

\*setimate for u + a3 + tau2 = 43 + 0 + 0 the same as using the beta hat estimates.