Math 416 – Exam 2
Spring 2009 – Marian Frazier

Name SOLUTIONS

Solve all problems, and be careful not to spend too much time on a particular problem. All necessary SAS files are in our usual folder (P:\data\math\Frazier\Regression). You may only use the SAS files mentioned below. **You may not alter these SAS files in any way.** You may also use a calculator (or a calculator on the computer, like Maple). **You may not use any other computer application**, including Minitab and Excel. The exam is worth a total of 205 points; point values for each part are in parentheses. To receive maximum credit, show all of your work. Good luck!

1. (80 pts) Is brain weight (in grams) associated with gestation period (in days) and/or litter size after accounting for the effect of body weight (in kg)? The data are provided in mammal.xls. Use the SAS program **Mammal.sas** to help you answer this question.

PART I: The first-order model

a. Specify the appropriate first-order regression model for this research question. (The theoretical model – not the estimated regression function.) (5)

b. Would you classify any of the Y observations as outliers? Explain. (5)

c. Would you classify any of the X observations as outliers? Explain. (5)

d. Estimate the parameters in the regression model you provided in part (a). (5)

e. Report and interpret the coefficient of multiple determination. (5)

f. Suppose that a model includes only body weight and litter size. What percent of the additional unexplained variation in brain weight is accounted for by adding gestation period to the model? (5)

g. Do you think that logarithms should be used in this situation? Explain. (10)

PART II: Answer the next several parts using the model

$$E\{lbrain\} = \beta_0 + \beta_1 lbody + \beta_2 lgest + \beta_3 llitter .$$

h. Comment on the appropriateness of this new model. Is it superior to the one in parts (a) and (d)? Be specific and make sure to discuss fit, model assumptions, etc. (20)

i. The researchers decided to study a fifth deer mouse. This mouse had a body weight of 0.024 kg, a gestation period of 24 days, and a litter size of 5. Using the new model, calculate and interpret a 99% prediction interval for the brain size of this mouse. (10)

j. Test the hypothesis $H_0 : \beta_2 \text{ and } \beta_3 = 0$. Make sure you interpret your results in terms of the problem. (10)

#1. Mammal.sas

Part I

a. $(\text{brain})_i = \beta_0 + \beta_1(\text{gestation})_i + \beta_2(\text{litter})_i + \beta_3(\text{weight})_i + \varepsilon_i$

b. Yes - there is a "brain" value at 5000 that is definitely an outlier, and two other values at ~800 and ~1600 that may be outliers.

c. gestation ≈ 650 days (1 value)
litter size > 6 (4 values)
body weight > 1000 kg (2 values)
} All of these lie far from the bulk of the data and thus, are possible outliers.

d. $b_0 = -225.292$
$b_1 = 0.9859$ where X_1 is weight
$b_2 = 1.8087$ where X_2 is gestation
$b_3 = 27.6486$ where X_3 is litter size

e. $R^2 = 0.8100$ 81% of the variation in brain weight is explained by the linear relation with {body weight, gestation, litter size} together.

f. $R^2_{Y2|13} = 0.22061$

g. Definitely some kind of transformation is necessary. The data is so "clumpy" down in the lower values, taking the natural log should spread things out and help reduce the extremity of those outliers.
Note that the residual plots are awful!

#1 Part II

h. The relationships between ln(brain) + ln(gestation) and ln(brain) + ln(litter) seem linear. Thus, it seems like these transformations are appropriate to apply a linear model to.
In addition, once you actually run the model, we see that the residual plots are much better. There are still some outlying residuals, but overall they show random scatter. The residuals are closer to normality, too. Overall, the model assumptions seem to be satisfied. (This was not true about the original, untransformed, model.)
Finally, R^2 of this model is much higher (95.37%), indicating a better fit.

i. Look at observation #35
Notice that Deer mouse IV has the same X_1, X_2, X_3 values as the "new" mouse.

$$\hat{Y}'_h = -0.46082$$

99% PI for $Y'_{h(new)} = \ln(\text{brain})_{h(new)}$:

$$-1.724426 \le Y'_{h(new)} \le 0.82262$$

$\Rightarrow$ 99% PI for $Y_{h(new)} = (\text{brain})_{h(new)}$:

$$e^{-1.724426} = 0.17477 \le Y_{h(new)} \le e^{0.82262} = 2.2765$$

For a deer mouse with these qualities, we are 99% confident that the brain size will be between 0.175 g and 2.277 g.

j. $H_0: \beta_2 = \beta_3 = 0$ vs. $H_a: \beta_2 \neq 0$ or $\beta_3 \neq 0$

$$F^* = \frac{SSR(X_2, X_3 | X_1)}{2} \div MSE$$

$$= \frac{SSR(X_2|X_1) + SSR(X_3|X_1, X_2)}{2} \div MSE$$

$$= \frac{9.062626 + 1.161247}{2} \div 0.22539$$

$$= 23.68$$

p-val $= P(F^*_{2,92} > 23.68) < 0.001$

Reject H_0; conclude that at least one of ln(gest) + ln(litter) contribute significantly to the model.

2. (45 pts) In a regression analysis of flower growth for different flower families, Y is a measure of flower height, X_1 is an measure of the brightness of the light available to each flower, and $X_2 - X_4$ are indicator variables for the flower family, coded as follows:

Type of flower	X_2	X_3	X_4
Oxalidaceae	1	0	0
Primulaceae	0	1	0
Apocynaceae	0	0	1
Scrophulariaceae	0	0	0

The response function to be used in the study is $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$.

a. Identify the response function for each type of flower. (16)

b. Interpret the following coefficients: (4 each)

   1) β_1
   2) β_3

c. For each of the following questions, specify the hypotheses H_0 and H_1 for the appropriate test: (4 each)

   1) With X_1 fixed, is the expected flower height the same for *Oxalidaceae* as *Apocynaceae*?

   2) With X_1 fixed, is the expected flower height the same for all four families?

d. With X_1 fixed, the four response functions in part (a) all have the same slope. Specify a more general response function that would allow each type of flower to have a different slope and identify the slope for each type of flower. (13)

#2. Flower growth

Sc.: $E\{Y\} = \beta_0 + \beta_1 X_1$

b. 1) β_1 is the slope for all types of flowers.
So if you increase the light by one unit, the height is expected to increase by β_1 units, regardless of flower type.

a) β_3 is the change in height for Pr., in comparison to Sc.

c. 1) $H_0: \beta_0 + \beta_2 = \beta_0 + \beta_4$ vs. $H_1: \beta_2 \neq \beta_4$

$\beta_2 = \beta_4$

2) $H_0: \beta_0 = \beta_0 + \beta_2 = \beta_0 + \beta_3 = \beta_0 + \beta_4$ vs. H_1: at least one $\beta_k \neq 0$ $(k = 2,3,4)$

$\beta_2 = \beta_3 = \beta_4 = 0$

d. Include interaction terms:

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_1 X_2 + \beta_6 X_1 X_3 + \beta_7 X_1 X_4$$

With this model,

Ox: slope $= \beta_1 + \beta_5$

Pr: slope $= \beta_1 + \beta_6$

Ap: slope $= \beta_1 + \beta_7$

Sc: slope $= \beta_1$

3. (80 pts) Healthy bones are continually being renewed by two processes. Through bone formation, new bone is built; through bone resorption, old bone is removed. If one or both of these processes is disturbed, by disease, aging, or space travel, for example, bone loss can result. The variables VOPLUS and VOMINUS measure bone formation and bone resorption, respectively. Osteocalcin (OC) is a biochemical marker for bone formation: higher levels of bone formation are associated with higher levels of OC. A blood sample is used to measure OC, and it is much less expensive to obtain than direct measures of bone formation. The units are milligrams of OC per millimeter of blood (mg/ml). Similarly, tartrate resistant acid phospahtase (TRAP) is a biochemical marker for bone resorption that is also measured in blood. It is measured in units per liter (U/l). These variables were measured for 31 healthy women aged 11 to 32 and are provided in the file biomarkers.dat. Use a scatter plot matrix (attached) and the SAS program **biomarkers.sas** to answer the questions below regarding the prediction of VOPLUS, the measure of bone formation.

PART I: The analyst, Brian, decided to fit the regression model with all three predictor variables.

a. Identify the estimated regression function for this model and comment on the fit. Would you suggest any changes to this model? (15)

PART II: Marian is another analyst. She suggested the regression model with all three predictor variables, the squared predictor variables, and all possible pairwise interaction terms.

b. Do you prefer this second-order model over the model in part (a)? Explain. (10)

c. Does there appear to be any problem with multicollinearity? Explain. (10)

PART III: Yet another analyst, Brad, decided to use the second-order model with the centered variables.

d. Looking at all the output, what model do you feel is the best? Justify your selection. (20)

e. Brad chose the (partial) second-order model using the following centered variables:

```
mvom moc mvomtrap mocsq mtrapsq
```

Derive expressions for the estimated regression coefficients (b0, b1, etc) for these five variables in terms of the *uncentered* variables. (That is, Brad has the estimated coefficients **b** using the centered variables. Derive expressions for the uncentered variables' coefficients **b'** in terms of **b**.) (25)

#3. Biomarkers.sas

Part I

a. Model: $\hat{y} = 243.488 + 0.9745 \cdot vom + 8.235 \cdot oc + 6.607 \cdot trap$

There are a few slightly outlying residuals, but other than that the residual plots look good. The normal prob. plot shows a normal assumption is reasonable for the error terms.

$R^2 = 0.8844$, which is pretty high.

Changes: It looks like TRAP can be dropped from the model - its coefficient is not significant.

Part II

b. The second-order model doesn't seem to be preferable to the first-order model. R^2_{adj} has increased, but only by 2%. None of the additional (second-order) variables seem to be significant. The outliers mentioned above are still there.

c. Yes! Not surprisingly, all of the variables are very highly correlated with their higher-order terms. She should center to avoid this problem.

However, centering will not change the correlation between the three original variables. The relationship between vom * oc ($r = .455$), and trap * oc ($r = .723$), and trap * vom ($r = .678$) are all large enough to bring up worries about multicollinearity, even in the first-order model.

d. My preferred model:

$$(rolled)_i = \beta_0 + \beta_1 (mvom)_i + \beta_2 (moc)_i + \beta_3 (mvom \times moc)_i + \beta_4 (mvom_i)^2 + \beta_5 (moc_i)^2$$

"Best" model that is both simple and includes all lower-order versions of higher-order terms.

#3

Part III

[redacted]

e. $\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_1X_3 + b_4X_2^2 + b_5X_3^2$

$= b_0 + b_1(X_1 - \bar{X}_1) + b_2(X_2 - \bar{X}_2) + b_3(X_1 - \bar{X}_1)(X_3 - \bar{X}_3)$
$+ b_4(X_2 - \bar{X}_2)^2 + b_5(X_3 - \bar{X}_3)^2$

$= b_0 + b_1X_1 - b_1\bar{X}_1 + b_2X_2 - b_2\bar{X}_2 + b_3(X_1X_3 - \bar{X}_1X_3 - X_1\bar{X}_3 + \bar{X}_1\bar{X}_3)$
$+ b_4(X_2^2 - 2X_2\bar{X}_2 + \bar{X}_2^2) + b_5(X_3^2 - 2X_3\bar{X}_3 + \bar{X}_3^2)$

$= b_0 + b_1X_1 - b_1\bar{X}_1 + b_2X_2 - b_2\bar{X}_2 + b_3X_1X_3 - b_3\bar{X}_1X_3 - b_3X_1\bar{X}_3 + b_3\bar{X}_1\bar{X}_3$
$+ b_4X_2^2 - 2b_4X_2\bar{X}_2 + b_4\bar{X}_2^2 + b_5X_3^2 - 2b_5X_3\bar{X}_3 + b_5\bar{X}_3^2$

$= (b_0 - b_1\bar{X}_1 - b_2\bar{X}_2 + b_3\bar{X}_1\bar{X}_3 + b_4\bar{X}_2^2 + b_5\bar{X}_3^2)$
$+ (b_1 - b_3\bar{X}_3)X_1 + (b_2 - 2b_4\bar{X}_2)X_2 + (-b_3\bar{X}_1 - 2b_5\bar{X}_3)X_3$
$+ b_3X_1X_3 + b_4X_2^2 + b_5X_3^2$

$b_0' = b_0 - b_1\bar{X}_1 - b_2\bar{X}_2 + b_3\bar{X}_1\bar{X}_3 + b_4\bar{X}_2^2 + b_5\bar{X}_3^2$

$b_1' = b_1 - b_3\bar{X}_3$

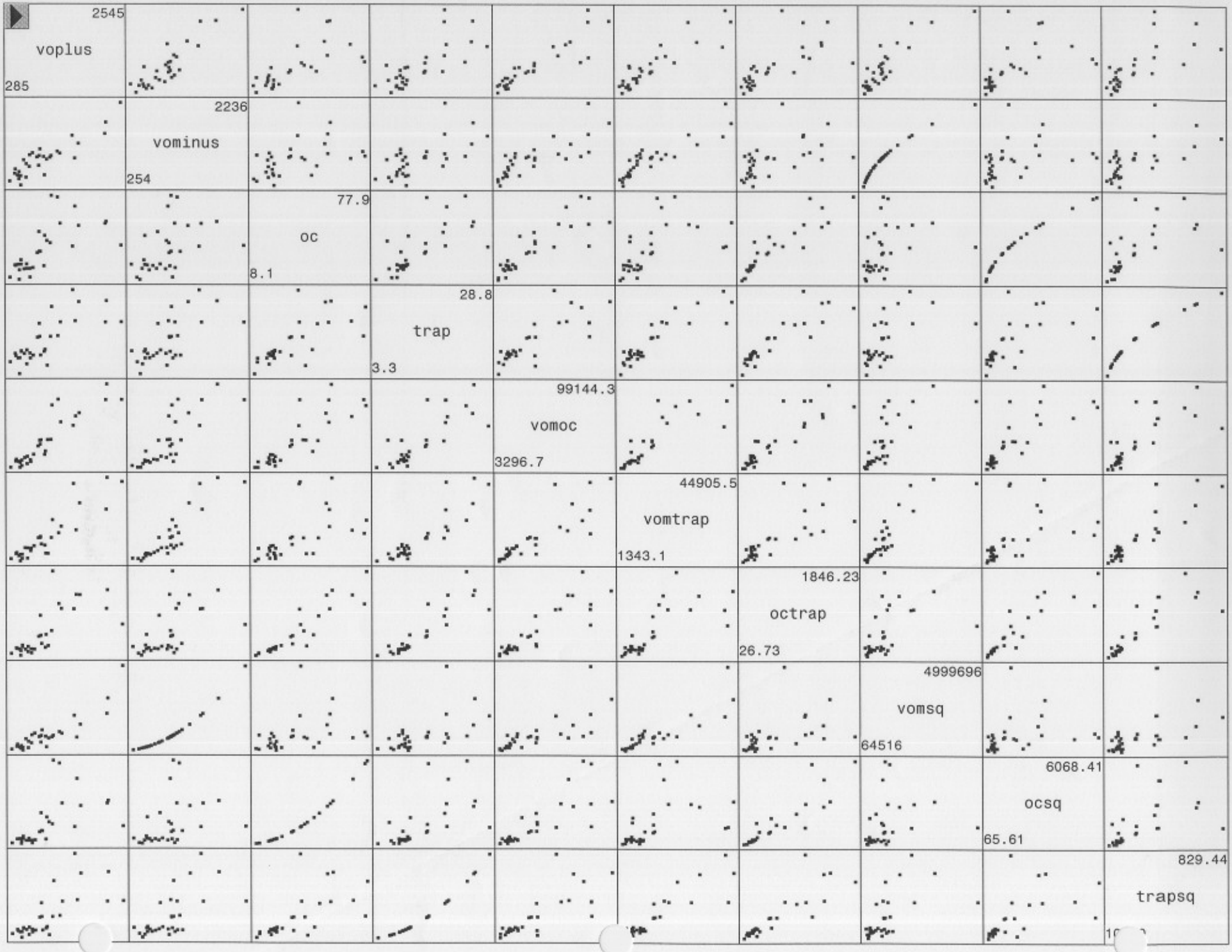
$b_2' = b_2 - 2b_4\bar{X}_2$

$b_3' = -b_3\bar{X}_1 - 2b_5\bar{X}_3$ ← coefficient on X_3

$b_3'' = b_3$ ← coefficient on X_1X_3

$b_4' = b_4$

$b_5' = b_5$


2545
voplus
285
2236
vominus
254
77.9
oc
8.1
28.8
trap
3.3
99144.3
vomoc
3296.7
44905.5
vomtrap
1343.1
1846.23
octrap
26.73
4999696
vomsq
64516
6068.41
ocsq
65.61
829.44
trapsq