Math 416 – Exam 1 Name _____SOLUTIONS_____
Spring 2009 – Marian Frazier

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 (embryos.sas and beetle.sas). **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 125 points; point values for each part are in parentheses. To receive maximum credit, show all of your work. Good luck!

1. (46 pts) The file P:\data\math\Frazier\Regression\embryos.dat gives the dry weights (Y) of 11 chick embryos ranging in age from 6 to 16 days (X). Use the file P:\data\math\Frazier\Regression\embryos.sas to answer the questions below.

   a. (6 pts) The values of the common logarithm of the weights (Z) are created and two scatterplots are provided. Describe the relationships between age (X) and dry weight (Y), and between age and $\log_{10}(dry\ weight)$.

   Weight (Y) vs. age (X) has a curvilinear pattern; a simple linear model would probably not be appropriate here.
   Using the transformed weight (Z) greatly improves the linear form; in fact, the relationship looks almost perfectly linear.

   b. (6 pts) State the simple linear regression models for these two regressions: Y regressed on X, and Z regressed on X.

$$\hat{Y} = -1.88453 + 0.23507X$$
$$\hat{Z} = -2.68920 + 0.19588X$$

   c. (6 pts) Which of the two regression lines in part (b) has a better fit? That is, is it more appropriate to run a linear regression of Y on X, or of Z on X? Explain your choice thoroughly.

   Z on X is clearly the better fit. It has a higher R^2 (99.83% vs. 74%). Plus, the standardized residual plot is much better – it looks like a random cloud of points, while the Y-on-X residual plot shows a bad curved pattern; not a surprise, considering what we saw in (a).

For the next three parts, use the regression that you chose as being more appropriate in part (c).

d. (12 pts) Find 95% confidence intervals for the true slope and intercept. Interpret each interval with regard to the null hypothesis that the true parameter is 0.

95% CI for β_0: (-2.7583, -2.62009)

We are 95% confident that the true value of the intercept is between these two values. This leads us to conclude that the intercept is significantly different from 0.

95% CI for β_1: (0.18984, 0.20192)

We are 95% confident that the true value of the slope is between these two values. This leads us to conclude that the slope is significantly different from 0. Thus, there is a significant linear relationship between age and log(weight).

e. (10 pts) Find and interpret joint confidence intervals for both the slope and intercept parameters with an overall (family) confidence coefficient of 0.98.

Joint (Bonferroni) 98% CIs: $\begin{matrix} b_0 \pm B \cdot s\{b_0\} \\ b_1 \pm B \cdot s\{b_1\} \end{matrix}$, where $B = t(1-\alpha/4; n-2) = t(1-0.02/4; 9)$
$= t(0.995; 9) = 3.25$

CI for β_0: $b_0 \pm B \cdot s\{b_0\} = -2.689 \pm 3.25 \cdot 0.03055 = (-2.7883, -2.5897)$

CI for β_1: $b_1 \pm B \cdot s\{b_1\} = 0.19588 \pm 3.25 \cdot 0.00267 = (0.18720, 0.20456)$

We are 98% confident that the true intercept lies between -2.79 and -2.59 **and** that the true slope lies between 0.187 and 0.205.

f. (6 pts) Find and interpret an approximate 95% confidence interval on the mean response for an 8-day-old chick.

$X_h = 8 \Rightarrow \hat{Z}_h = -1.122$ This corresponds to observation 3 in the data set.

As can be seen in the SAS output below, the 95% CI for $E\{Z_h\}$ is (-1.1485, -1.0958). Thus, we are 95% confident that for chicks 8 days old, the true mean value of log(weight) is between -1.1485 and -1.0958. (Or, the true mean value of weight is 0.0710 and 0.0802.)

The REG Procedure
Model: MODEL1
Dependent Variable: Z

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	-1.5376	-1.5139	0.0158	-1.5496	-1.4782	-1.5866	-1.4412	-0.0237
2	-1.2840	-1.3180	0.0136	-1.3488	-1.2872	-1.3885	-1.2476	0.0340
3	-1.1024	-1.1222	0.0116	-1.1485	-1.0958	-1.1907	-1.0536	0.0198
4	-0.9031	-0.9263	0.009988	-0.9489	-0.9037	-0.9935	-0.8590	0.0232
5	-0.7423	-0.7304	0.008853	-0.7504	-0.7104	-0.7968	-0.6640	-0.0119

2. (48 pts) In a study on geographic variation in a certain species of beetle, the mean tibia length (U) and the mean tarsus length (V) were obtained for samples of size 50 from each of 10 different regions spanning five southern states. The results are provided in P:\data\math\Frazier\Regression\beetle.dat. Use the SAS program P:\data\math\Frazier\Regression\beetle.sas to answer the questions below.

   a. (4 pts) Find the estimated least squares equation for predicting tibia length (in mm) from tarsus length (in mm).

   We want to predict U from V, which means V is the predictor variable and U the response.
   Fitted regression line: $\hat{U} = 0.66072 + 4.06912V$

   b. (6 pts) Evaluate the fit of the model by looking at the residual plots.

   Looking at the scatterplot of U vs. V, a linear model seems to be appropriate. The standardized residual plots (sresidU vs. yhatU and V) show no outliers; however, there may be some non-constancy of error variance (megaphone shape?).

   c. (6 pts) Do you think the normality assumption is reasonable in this situation? Justify your response.

   The normal probably plot of residU doesn't look very straight, although it is hard to tell with so few data points. The stem-and-leaf and boxplots (in Proc Univariate) show a little left-skewness, but probably not enough to rule out normality. The tests for normality (Anderson-Darling, Cramer-von-Mises, etc) have very large p-values, indicating that there is no evidence to reject the assumption of normality of the error terms.

   d. (6 pts) Find and interpret an approximate 98% prediction interval on the response for a beetle• with tarsus length of 1.776 mm.

   $V_h = 1.776 \quad \Rightarrow \quad \hat{U}_h = 7.8875$ This corresponds to observation 7 in the data set.
   As can be seen in the SAS output below, the 98% PI for $U_{h(new)}$ is (7.5999, 8.1751). Thus, we are 98% confident that a given beetle with tarsus length of 1.776 mm will have a tibia length of between 7.5999 mm and 8.1751 mm.

The REG Procedure
Model: MODEL1
Dependent Variable: U

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	98% CL Mean		98% CL Predict		Re
1	7.5000	7.4968	0.0329	7.4016	7.5921	7.2099	7.7838	0.
2	7.1640	7.1550	0.0505	7.0088	7.3013	6.8474	7.4627	0.
3	7.5120	7.4968	0.0329	7.4016	7.5921	7.2099	7.7838	
4	8.5440	8.4246	0.0648	8.2370	8.6122	8.0953	8.7539	
5	7.3800	7.3015	0.0418	7.1804	7.4226	7.0050	7.5980	
6	7.8600	7.7898	0.0307	7.7009	7.8788	7.5049	8.0747	
7	7.8360	7.8875	0.0336	7.7903	7.9847	7.5999	8.1751	-
8	8.1000	8.2293	0.0517	8.0795	8.3791	7.9199	8.5387	-

• This question should actually read "Find …PI on the mean tibia length for a group of beetles, all with tarsus lengths of 1.776 mm."

e. (10 pts) Using $\alpha = .05$, conduct a formal test for lack of fit.

$$H_0 : E\{Y\} = \beta_0 + \beta_1 X \qquad H_a : E\{Y\} \neq \beta_0 + \beta_1 X$$

From Proc Reg output, $SSE = 0.06985$. From Proc GLM, $SSPE = 0.017568$ with 2 d.f.
So $SSLF = 0.06985 - 0.017568 = 0.05228$ with 6 d.f.
Thus the test statistic is

$$F^* = \frac{MSLF}{MSPE} = \frac{0.05228/6}{0.017568/2} = \frac{0.008713}{0.008784} = 0.9919$$

Comparing this to an F(6,2) distribution, we would reject H0 for F* > 19.3. Put another way, p-value = $P(F_{6,2} > F^*) > 0.5$. So obviously, we would not reject H0; thus, we conclude that the linear model seems to be appropriate for this data.

f. (8 pts) Report the appropriate sample correlation coefficient between tarsus length and tibia length. Explain why you choose that correlation coefficient.

Pearson correlation coefficient: $r_{12} = 0.976$
Spearman correlation coefficient: $r_S = 0.951$

Using the Pearson measure is only valid if U and V are jointly bivariate normal. In order to test if this is true, we can look at the normality of each variable individually. Stem-and-leaf and boxplots for both U and V show that both variables are a bit right-skewed. However, tests for normality for both variables are not significant at any acceptable level, indicating that there is no reason to reject the assumption of normality.
There is so little data that it is difficult to make a firm conclusion about normality. I think an argument could be made for either measure of correlation here.

g. (8 pts) Using the statistic found in part (f), test the hypotheses $H_0 : \rho = 0$ versus $H_A : \rho \neq 0$. Make sure to report the test statistic, the p-value, and a thorough conclusion.

Using Pearson:

Test statistic: $t^* = \frac{0.976\sqrt{10-2}}{\sqrt{1-0.976^2}} = 12.88$

p-value = $P(|t_8| > 12.88) < 0.0001$

Using Spearman:

Test statistic: $t^* = \frac{0.951\sqrt{10-2}}{\sqrt{1-0.951^2}} = 8.70$

p-value = $P(|t_8| > 8.70) < 0.0001$

So regardless of which statistic is used, we reject the null hypothesis and conclude that there is a significant correlation between mean tarsus length and mean tibia length.

3. (31 pts) Suppose that in the model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, the errors have mean zero and are independent, but $Var(\varepsilon_i) = \kappa_i^2 \sigma^2$, where the κ_i are known constants, so the errors do not have equal variance. This situation arises when the Y_i are averages of several observations at X_i; in this case, if Y_i is an average of n_i independent observations, $\kappa_i^2 = \frac{1}{n_i}$.

   a. (12 pts) The model may be transformed as follows: $\kappa_i^{-1} Y_i = \kappa_i^{-1}\beta_0 + \kappa_i^{-1}\beta_1 X_i + \kappa_i^{-1}\varepsilon_i$ or $Z_i = \beta_0 U_i + \beta_1 V_i + \gamma_i$ where $U_i = \kappa_i^{-1}, V_i = \kappa_i^{-1} X_i, \gamma_i = \kappa_i^{-1}\varepsilon_i$.
   Show that the new model satisfies the assumptions of the standard linear regression model.

   <u>Assumptions:</u>

   1. Expected value of the error terms is 0.
   $$E\{\gamma_i\} = E\{\kappa_i^{-1}\varepsilon_i\} = \kappa_i^{-1} E\{\varepsilon_i\} = \kappa_i^{-1}(0) = 0 \qquad \text{☺}$$

   2. Variance of the error terms is constant over the predictor (X).
   $$Var\{\gamma_i\} = Var\{\kappa_i^{-1}\varepsilon_i\} = \kappa_i^{-2} Var\{\varepsilon_i\} = \kappa_i^{-2}\kappa_i^2\sigma^2 = \sigma^2 \qquad \text{☺}$$

   3. Error terms are independent of each other.
   $$Cov\{\gamma_i, \gamma_j\} = Cov\{\kappa_i^{-1}\varepsilon_i, \kappa_j^{-1}\varepsilon_j\} = Cov\{\varepsilon_i, \varepsilon_j\}, \text{ since the } \kappa_i \text{ are constants}$$
   $$= 0, \text{ since the } \varepsilon_i \text{ are independent} \qquad \text{☺}$$

   4. Values of predictor (X) can be thought of as constants.
   Here, we have two predictors: U_i and V_i. Since $U_i = \kappa_i^{-1}$ and the κ_i are known constants, it follows that U_i are constants, as well. And if we assume that the values of the original predictor variable X_i are constants, it must be that $V_i = \kappa_i^{-1} X_i$ are constants.

b. (10 pts) Using the model for Z_i in part (a), identify the normal equations for finding the least squares estimators of β_0 and β_1. DO NOT solve these simultaneous equations.

The least squares criterion for this model is

$$Q = \sum_{i=1}^{n} \left(Z_i - \beta_0 U_i - \beta_1 V_i \right)^2$$

We wish to find b_0 and b_1 that minimize Q.

Take the partial derivative of Q with respect to β_0 and β_1:

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^{n} U_i \left(Z_i - \beta_0 U_i - \beta_1 V_i \right)$$

$$\frac{\partial Q}{\partial \beta_1} = -2 \sum_{i=1}^{n} V_i \left(Z_i - \beta_0 U_i - \beta_1 V_i \right)$$

When we set these partials equal to 0, that will give us b_0 and b_1:

$$-2 \sum_{i=1}^{n} U_i \left(Z_i - b_0 U_i - b_1 V_i \right) = 0 \qquad -2 \sum_{i=1}^{n} V_i \left(Z_i - b_0 U_i - b_1 V_i \right) = 0$$

$$\text{(1)} \quad \sum U_i Z_i - b_0 \sum U_i^2 - b_1 \sum U_i V_i = 0 \qquad \sum V_i Z_i - b_0 \sum V_i U_i - b_1 \sum V_i^2 = 0 \quad \text{(2)}$$

Equations (1) and (2) are the normal equations for this model. Solving this system will result in formulas for the least squares estimators b_0 and b_1.

c. (9 pts) Show that performing a least squares analysis on the new model, as was done in part (b), is equivalent to minimizing

$$\sum_{i=1}^{n}\left(Y_i-\beta_0-\beta_1X_i\right)^2\kappa_i^{-2}\ .$$

If we want to minimize $Q_2=\sum_{i=1}^{n}\left(Y_i-\beta_0-\beta_1X_i\right)^2\kappa_i^{-2}$, we must take the partial derivatives with respect to β_0 and β_1, and set those partials equal to 0 in order to find b_0 and b_1, just as we did in part (b).

$$\frac{\partial Q_2}{\partial\beta_0}=-2\sum_{i=1}^{n}\kappa_i^{-2}\left(Y_i-\beta_0-\beta_1X_i\right)\qquad\frac{\partial Q_2}{\partial\beta_1}=-2\sum_{i=1}^{n}\kappa_i^{-2}X_i\left(Y_i-\beta_0-\beta_1X_i\right)$$

$$(3)\qquad\sum_{i=1}^{n}\kappa_i^{-2}\left(Y_i-b_0-b_1X_i\right)=0\qquad\sum_{i=1}^{n}\kappa_i^{-2}X_i\left(Y_i-b_0-b_1X_i\right)=0\qquad(4)$$

We want to show that Equations (1) and (2) are equivalent to Equations (3) and (4).

Consider Equation (1) above; plug in the definitions of Z_i, U_i, and V_i:

$$(1)\qquad\begin{aligned}\sum U_iZ_i-b_0\sum U_i^2-b_1\sum U_iV_i&=0\\ \sum\kappa_i^{-1}\kappa_i^{-1}Y_i-b_0\sum\kappa_i^{-2}-b_1\sum\kappa_i^{-1}\kappa_i^{-1}X_i&=0\end{aligned}$$

$$(5)\qquad\sum_{i=1}^{n}\kappa_i^{-2}\left(Y_i-b_0-b_1X_i\right)=0$$

Similarly for Equation (2):

$$(2)\qquad\begin{aligned}\sum V_iZ_i-b_0\sum V_iU_i-b_1\sum V_i^2&=0\\ \sum\kappa_i^{-1}X_i\kappa_i^{-1}Y_i-b_0\sum\kappa_i^{-1}X_i\kappa_i^{-1}-b_1\sum\kappa_i^{-1}X_i\kappa_i^{-1}X_i&=0\end{aligned}$$

$$(6)\qquad\sum_{i=1}^{n}\kappa_i^{-2}X_i\left(Y_i-b_0-b_1X_i\right)=0$$

Notice that Equations (3) and (5) are the same; so are Equations (4) and (6). Thus, the analysis we did in part (b) is equivalent to minimizing Q_2.