Math 484 Exam 2 Fall 2016 Name___________

1) Suppose a researcher desires to predict average female score from average male score from data on standardized science tests given to 8th graders in 39 countries. Assume that the correlation $r = 0.9889$. $= \hat{\rho},$

predict Y from X

	variable	mean	standard deviation
Y	female score	508.79	49.304
X	male score	525.38	49.525

a) Find the slope of the least square line.

$$\hat{\beta}_2 = \hat{\rho}\,\frac{s_Y}{s_X} = 0.9889\,\frac{49.304}{49.525} = \boxed{0.9845}$$

b) Find the intercept of the least square line.

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} = 508.79 - 0.9845(525.38)$$

$$= \boxed{-8.4466}$$

14

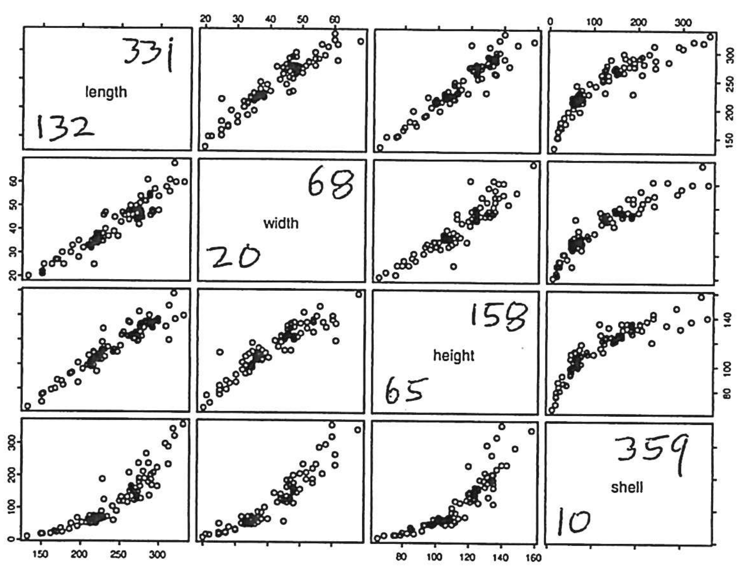

Figure 1: Scatterplot Matrix for Original Mussel Data Predictors

2) From the figure above, which transformations are suggested by the log rule?

$\boxed{\log(\text{shell})}$ since $\frac{359}{10} > 10$

7

3) From the figure below, which transformation should be used? Explain briefly.

$y = \log(z)$ since the plot is linear

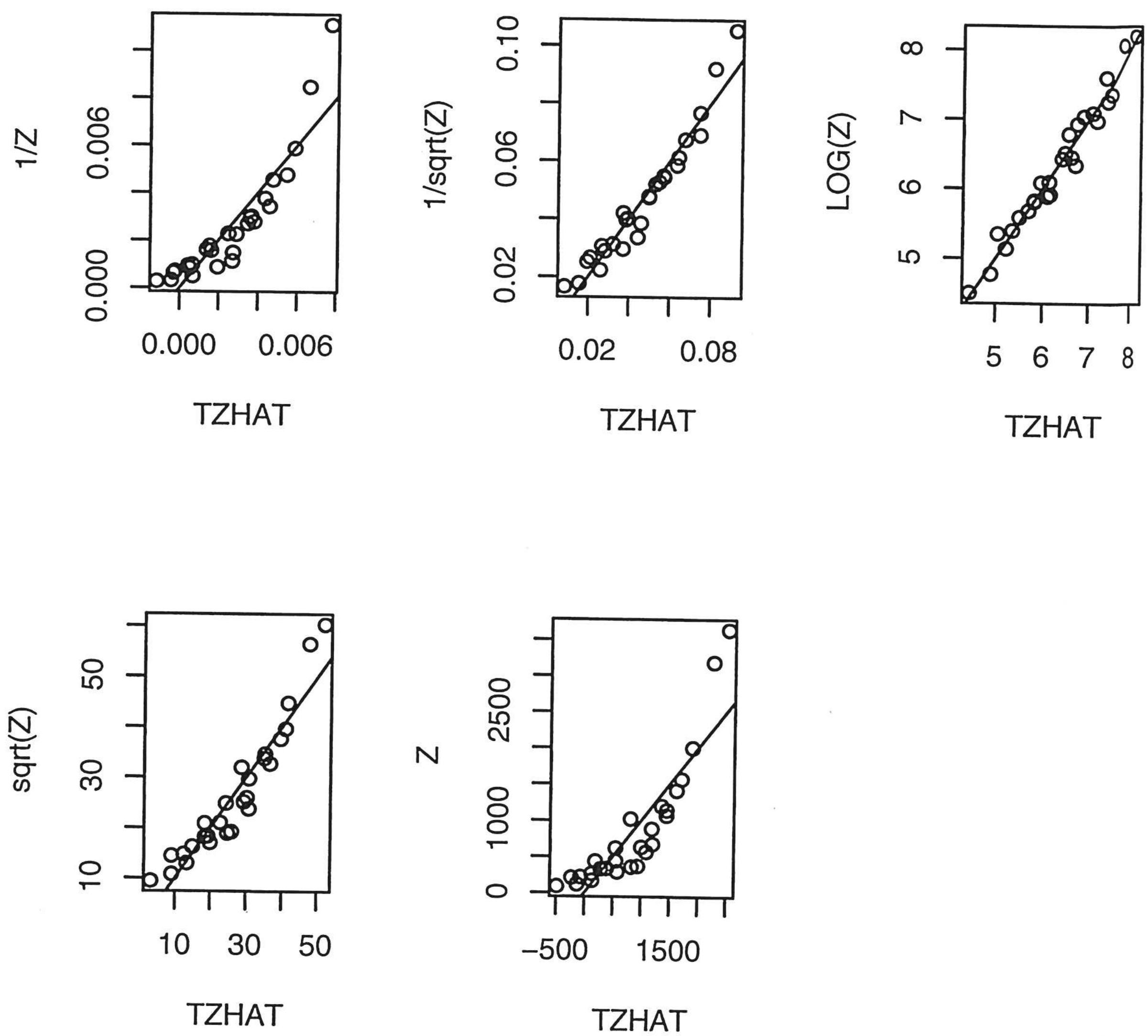

Figure 1:

7

Math 484 Exam 2 Fall 2016

	Estimate	Std.Err	t-value	Pr(>|t|)
X	0.9966	0.0016	626.6118	0

X = remembered score

Summary Analysis of Variance Table

Source	df	SS	MS	F	p-value
Regression	1			39264.2	0.0000
Residual	19				

4) Dr. Olive graded 2016 quiz 5, then entered the scores on the computer. He tried to remember the score x_2 and then entered the actual score Y for the 20 Math 484 students. Regression through the origin was used. Assume least squares output can be used.

a) Predict Y if $x_2 = 90$. $\hat{Y} = \hat{\beta}_1 x = 0.9966(90) = \boxed{89.694}$

b) Perform the t test for $\beta_2 = 0$.

Ho $\beta_1 = 0$ HA $\beta_1 \neq 0$

$t_0 = 626.6118$

pval = 0

reject Ho X = remembered score is needed in the MLR model for Y = actual score

c) Perform the no intercept Anova F test.

Ho $\beta_1 = 0$ HA $\beta_1 \neq 0$

$F_0 = 39264.2$

pval = 0

reject Ho there is an MLR relationship between Y = actual score and X = remembered score

21

$-2 \quad \overline{\underline{2}}$ 34→30 2.042 (500.57, 524.11)

(439.90, 584.74)

```
Label        Estimate        Std. Error     t-value     p-value
Constant     445.862         26.3328        16.932      0.0000
edaids         0.787074       0.287856        2.734       0.0099
fteach         0.578163       0.330786        1.748       0.0895
Prediction = 512.341, se(pred) = 35.4772, se = 5.76336
```

Q4) 22

5) Let Y = *mean science score* of female 8th graders in 37 countries. The predictors are *edaids* = percent of 8th graders with dictionary, study table and computer and *fteach* = percentage of 8th graders taught by female teachers. Suppose that it is desired to predict Y_f if *edaids* = 47 and *fteach* = 51, so that $\boldsymbol{x}_f = (1, 47, 51)^T$. Assume that $\hat{Y}_f = 512.341$, $se(\hat{Y}_f) = 5.76336$ and that se(pred) = 35.4772.

0F -3

a) If $\boldsymbol{x}_f = (1, 47, 51)^T$ find a 95% confidence interval for $E(Y_f|\boldsymbol{x}_f)$.

$df = n-3 = 34 \to \infty$ 1.96 95%

$\hat{Y}_f \pm t_{n-p, 1-\frac{\delta}{2}} SE(\hat{Y}_f) = 512.341 \pm 1.96\,(5.76336)$

$= 512.341 \pm 11.296 = \boxed{[501.045, 523.637]}$

b) If $\boldsymbol{x}_f = (1, 47, 51)^T$ find a 95% prediction interval for Y_f.

$\hat{Y}_f \pm t_{n-p, 1-\frac{\delta}{2}} SE(pred) = 512.341 \pm 1.96\,(35.4772)$

$= 512.341 \pm 69.535 = \boxed{[442.806, 581.876]}$

12

	L1	L2	L3	L4
# of predictors	13	6	5	4
# with $0.01 \le$ p-value ≤ 0.05	0	0	0	0
# with p-value > 0.05	9	2	1	2
R_I^2	0.993	0.991	0.990	0.981
corr$(\hat{Y}, \hat{Y}_I)$	1.0	0.9990	0.9987	0.9948
$C_p(I)$	13.0	4.70	5.24	27.21
$\sqrt{MSE}$	4.024	3.933	4.040	5.322
p-value for partial F test	1.0	0.985	0.444	0.007

L2 = I_{min}

6) The above table gives summary statistics for 4 MLR models considered as final submodels after performing variable selection. The data set had 35 cases and assume that the full model L1 was good even though n is small. Which model should be used as the final submodel? Explain briefly why each of the other 3 submodels should not be used.

L3 $= I_I$ should be used $\quad 5.24 < 4.7 + 1 = 5.7$

L1 and L2 have too many predictors

L4 has $C_p > 2k$ and pval for partial F test is too small

6

7) Suppose that the regression model is $Y_i = \beta + 7X_i + \epsilon_i$ for $i = 1, \ldots, n$ where the ϵ_i are iid $N(0, \sigma^2)$ random variables. The least squares criterion is $Q(\eta) = \sum_{i=1}^{n} (Y_i - \eta - 7X_i)^2$.

a) What is $E(Y_i)$?

$$\boxed{\beta + 7X_i}$$

b) Find the least squares estimator $\hat{\beta}$ of β by setting the first derivative $\frac{d}{d\eta} Q(\eta)$ equal to zero.

$$\frac{d}{d\eta} Q(\eta) = -2 \sum (Y_i - \eta - 7X_i) \overset{set}{=} 0$$

$$\text{or} \quad \sum (Y_i - 7X_i) = n\eta$$

$$\text{so } \hat{\beta} = \boxed{\frac{\sum (Y_i - 7X_i)}{n} = \bar{Y} - 7\bar{X}}$$

c) Show that your $\hat{\beta}$ is the global minimizer of the least squares criterion Q by showing that the second derivative $\frac{d^2}{d\eta^2} Q(\eta) > 0$ for all values of η.

$$\frac{d^2}{d\eta^2} Q(\eta) = \frac{d}{d\eta} \left[-2 \sum (Y_i - \eta - 7X_i) \right]$$

$$= \frac{d}{d\eta} \left[-2 \sum (Y_i - 7X_i) + 2n\eta \right]$$

$$= 2n > 0$$

18

eleventh of 10 students

Current terms: (brate gnp idrate mlife)

	df	RSS	|	k	C_I
Delete: gnp	93	242.17	|	4	2.677
Delete: brate	93	280.647	|	4	17.243
Delete: idrate	93	315.003	|	4	30.248
Delete: mlife	93	735.697	|	4	189.508

8) The above output for Rouncefield (1995) data is for predicting Y = *female life expectancy* from x_2 = *brate* (birth rate), x_3 = *drate* (death rate), $x_4 = GNP$, x_5 = *idrate* (infant death rate), and x_6 = *mlife* (male life expectancy). What is the I_{min} model that hs the smallest $C_p(I)$? Do not forget the constant.

constant, brate, idrate, mlife

7

9) For the figure below, consider the ladder rule for making a transformation $t_\lambda(w)$ to increase the linearity of the plot.

a) If $w = Y$, should λ be made larger or smaller? smaller

b) If $w = X$, should λ be made larger or smaller? smaller

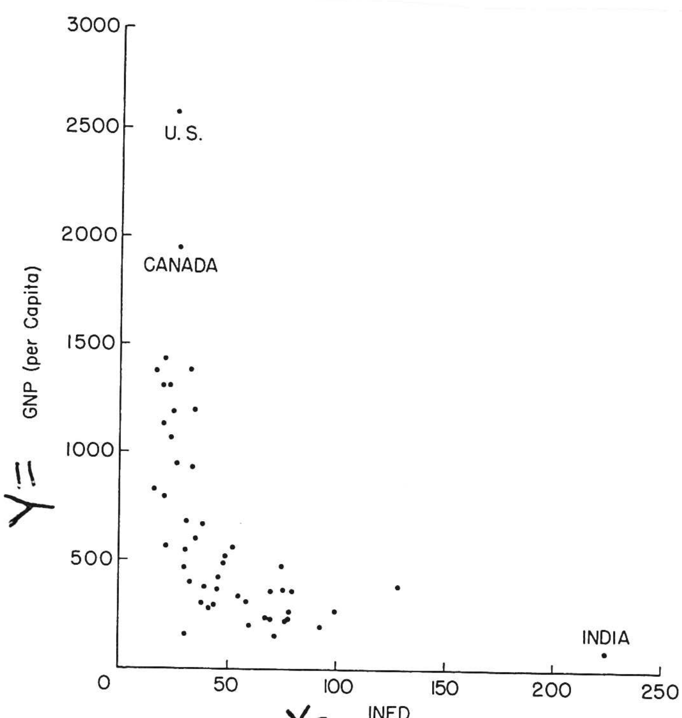

Figure 2.4 Gross national product (GNP) per capita by infant death rate (INFD) in 49 countries in the world.

Gunst and Mason (1980, p.31)

8