Math 356H Assignment #3

Readings

- (NWK) Sections 6.8 to 6.13 (skipping ANOVA for now). The matrix approach allows you to visualize all the horrendous formulas in a nice, clear form. If you like Linear Algebra, you will enjoy learning about expectation and variance of random matrices, as well as performing least squares with them.
- (D) Sections 13.3 and 13.4.
- *Optional*: (NWK) Sections 7.1 - 7.7. You can see the estimators in matrix form here. The complete worked out example is worth reading in detail.

Due date: Wednesday, February 27 (after Reading Week).

1. The results shown below were obtained in a small-scale experiment to study the relation between °F of storage temperature (X) and number of weeks before flavour deterioration of a food product begins to occur (Y).

i	1	2	3	4	5
X_i	8	4	0	-4	-8
Y_i	7.8	9.0	10.2	11.0	11.7

Assume that the first-order regression model is applicable. Using matrix methods (`R` is great for multiplying matrices!), find:

(a) the vector of estimated regression coefficients

(b) the vector of residuals

(c) the variance-covariance matrix of the vector of coefficients.

2. Consider the following Excel output, and use it to answer the given questions.

Regression Statistics	
Multiple R	0.409795
R square	0.167932
Adjusted R Square	0.145239
StandardError	1.067112
Observations	114

ANOVA

Source	df	SS	MS	F	Significance F
Regression	3	25.28057	8.426856	7.400232	0.000146209
Residual	110	125.2601	1.138728		
Total	113	150.5407			

	Coefficient	Standard Error	t Stat	P-value
Intercept	-2.02693	1.26949	-1.59665	0.113212
Latitude	0.069004	0.017405	3.964659	.000131
Longitude	0.008697	0.005985	1.453163	.149025
Depth	-0.02623	0.012521	-2.09508	.038923

(a) Construct the multiple regression equation that expresses earthquake magnitude (in ML) in terms of latitude (in degrees), longitude (in degrees) and depth (in meters).

(b) Test the overall significance of the multiple regression equation using $\alpha = .05$.

(c) Find the adjusted value of the coefficient of determination and interpret it.

(d) Is the multiple regression equation usable for predicting an earthquake's magnitude based on its recorded latitude, longitude, and depth? Explain briefly why or why not.

(e) Seismic activity has been detected at 48.2°N latitude, and 124.99°W longitude, at a depth of 1 m. Find the point estimate of the predicted magnitude of the earthquake. If the seismic activity in question actually had a magnitude of 0.9ML, find the residual.

3. Set up the X matrix and β vector for the following regression model (assume $i = 1, \ldots, 4$):

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1}^2 + \varepsilon_i.$$

4. A group of physicians hired a management consultant to see if the patients' waiting times could be reduced. The consultant randomly sampled 200 patients and found the average waiting time was 32 minutes, with a standard deviation of 15 minutes. To determine the factors that affected waiting time, the consultant fit the following multiple regression:

$$WAIT = 22 + .09DRLATE - .24PLATE + 2.61SHORT$$

where $WAIT$ is the waiting time, $DRLATE$ is the lateness of the doctors in arriving that morning (sum of their times), $PLATE$ was the lateness of the patient in arriving for their appointment and $SHORT$ was an indicator variable that equaled 1 if the clinic was short staffed, and equaled 0 if fully staffed with all 4 physicians. All times are in minutes.

The coefficient of determination was $R^2 = .72$ and the standard errors of the $DRLATE$, $PLATE$, and $SHORT$ regression coefficient estimates were .01, .05, and 1.38 respectively.

(a) Perform a model utility test for this model.

(b) If a patient is drawn at random, find a point estimate of the time he/she will have to wait:

i. If nothing else is known;

ii. If the patient is 20 minutes late, on a day when the clinic was fully staffed, but the four physicians were late by 10, 25, 15 and 20 minutes;

iii. If neither the patient nor the doctors are late and the clinic is fully staffed.

(c) What would you estimate as the difference in waiting time if, all other things being equal, the clinic is fully staffed as opposed to being short-staffed.

(d) Is the following statement TRUE or FALSE?

"Since the coefficient for $SHORT$ is the largest, it is the most important factor in accounting for the variation in $WAIT$." Explain your answer briefly.

5. A study of pregnant grey seals involved $n = 25$ observations on the variables y = fetus progesterone level (in milligrams), x_1 = fetus length (in centimetres), and x_3 = fetus weight (in grams). Part of the R output for the model using all three independent variable is given ("Gonadoterophin and Progesterone Concentration in Placenta of Grey Seals," *Journal of Reproduction and Fertility* (1984): 521-528):

Coefficients:

	Estimate	Std. Error	t-value
(Intercept)	-1.982	4.290	-0.46
X1	-1.871	1.709	-1.09
X2	0.2340	0.1906	1.23
X3	.000060	.002020	.03

Residual standard error: = 4.189

Multiple R-Squared: 0.552 Adjusted R-SquaredR-sq(adj) = 0.488

F statistic : 8.63 on 3 and 21 DF, p-value: .001

(a) Use information from the R output to test the hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$. (Use $\alpha = .05$.)

(b) Using an elimination criterion of $-2 \le t$ ratio ≤ 2, should any variable be eliminated? If so, which one?

(c) Part of the `R` output for the regression using only X_1 = sex and X_2 = length is given here:

Coefficients

	Estimate	Std. Error	t-ratio
(Intercept)	-2.090	2.212	-0.94
X1	-1.865	1.661	-1.12
X2	0.23952	0.04604	5.20

Residual std. error: 4.093

Multiple R-Squared: 0.552 Adjusted R-Squared: 0.512

Would you recommend keeping both X_1 and X_2 in the model? Explain.

(d) After elimination of both X_3 and X_1, the estimated regression equation is $\hat{Y} = -2.61 + .231X_2$. The corresponding values of R^2 and s are .527 and 4.116, respectively. Interpret these two values.

(e) Interpret the coefficients obtained in part (d).

6. Chapter 13, #44