Math 356H Assignment #3 Solutions

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

The X matrix is given by

$$X = \begin{bmatrix} 1 & 8 \\ 1 & 4 \\ 1 & 0 \\ 1 & -4 \\ 1 & -8 \end{bmatrix},$$

then the vector of estimated regression coefficients is

$$b = (X'X)^{-1}X'Y = \begin{bmatrix} 9.94 \\ -0.245 \end{bmatrix}.$$

(The `R` instruction would be `b=ginv(t(x)%*%x)%*%t(x)%*%y`.)

(b) the vector of residuals

With `e=y-x.matrix%*%b`, we get

$$\mathbf{e} = (-0.18, 0.04, 0.26, 0.08, -0.20)'$$

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

$s^2\{b\} = MSE(X'X)^{-1}$, with $MSE = e'e/(n-2)$. Hence

$$s^2\{b\} = \begin{bmatrix} .0148 & 0 \\ 0 & .0004625 \end{bmatrix}$$

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).

Let Y= earthquake magnitude then:

$EY = -2.03 + .069 LATITUDE + .009 LONGITUDE - .026 DEPTH$

Alternately, define x_1= latitude, x_2 = longitude and x_3= depth. Then

$EY = -2.03 + .069x_1 + .009x_2 - .026x_3$.

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

The P value for the model validation test is .0001, so at $\alpha = .05$, the null hypothesis of all coefficients being zero must be rejected. Hence there is at least one parameter which is non-zero.

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

The value of R^2 is .1679, so that 16.79% of variation in earthquake magnitude is explained by the three variables in the model. The adjusted value of R^2 is 14.52%, which represents the variation explained by the model, taking into consideration the additional variables incorporated (R^2 always increases, whereas r^2 adjusted does not).

(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.

Yes, but with caution. The equation can be used since the P-value for model validation is small. However, care must be exercised because not all coefficients are significantly different than zero, so perhaps a different model is more appropriate. Moreover, the R^2 is small, so that the amount of variation explained is not very high. Further analysis of residual plots is recommended to verify that the models assumptions are satisfied so that intervals, and not only point estimates, can be obtained.

(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.

$Y = -2.03 + .069(48.2) + .009(124.99) - .026(1) = 2.39471$.

Residual: 2.39471 - .9 = 1.49471.

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

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

$$X = \begin{pmatrix} 1 & X_{11} & X_{12} & X_{11}^2 \\ 1 & X_{21} & X_{22} & X_{21}^2 \\ 1 & X_{31} & X_{32} & X_{31}^2 \\ 1 & X_{41} & X_{42} & X_{41}^2 \end{pmatrix}$$

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:

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

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.

To test the hypothesis that all coefficients are zero, we use

$$F = \frac{R^2/k}{(1-R^2)/[n-(k+1)]} = \frac{.72/3}{(1-.72)/196} = 168.$$

This value is significant at $\alpha = .05$, compared to an $F_{.05,3,196}$, so we reject H_0 and there is some coefficient that is not zero.

(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;

We would estimate through $\bar{y} = 32$ minutes.

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;

$PLATE = 20, SHORT = 0, DRLATE = 70$. Since the model utility test has rejected no utility, we can use the regression line to make a point estimate of the average waiting time at this point. Hence the average waiting time is estimated to be

$WAIT = 22 + .09(70) - .24(20) + 2.61(0) = 23.5$ minutes.

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

In that case all variables are zero and the expected waiting time is 22 (the value of the intercept).

(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.

The difference would be the coefficient of $SHORT$, that is, 2.61 more if 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.

False. The variable $SHORT$ is in fact not influential at all, since a test for that coefficient fails to reject H_0 (the t-ratio for that coefficient is $2.61/1.38 = 1.89$, which is not significant at $\alpha = .05$ in a t-distribution with 196 degrees of freedom - we can use a normal distribution as an approximation).

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 $\alpha = .05$.)

This is equivalent to the model utility test, which from the ANOVA output can be seen to be significant at $\alpha = .05$.

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

Yes, x_3 has the smallest t-ratio, and hence whould be eliminated from the model.

(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.

No, the t-ratio for x_1 is still below the threshold of 2, so it should be eliminated from the model.

(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.

52.7% of variation is explained by the model. The estimate for σ, the standard deviation of errors, is 4.116.

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

The intercept cannot be interpreted since $x_2 = 0$ (a fetus of length 0) is not within the scope of the model. The value of .231 means that for every centimetre in increase in fetus length, there is an increase of .231 milligrams of progesterone.

6. Chapter 13, #44

(a) $\hat{\beta}_1 = .33563$ represents that for every percentage point of increase in flour protein there is an average increase of .33563 percentage points in absorption. Similarly, $\hat{\beta}_2 = 1.44228$ represents that for every Farrand unit of increase in starch damage, there is an increase of 2.44228 percentage points of absorption.

(b) $R^2 = .96447$, so that 96.446% of variation in absorption is explained by the model. Taking into account the inclusion of 2 variables, the R^2 adjusted is 96.16%.

(c) Yes. The p-value for model validity is .0000, so we reject the null hypothesis of all coefficients being 0 and there is a useful linear relationship between absorption and at least one of the two predictors.

(d) No. We carry a test for significance of the starch coefficient, keeping all other variables in the model by using the confidence interval provided. Since the CI does not include 0, the elimination of starch at 95% confidence is not justified.

(e) A 95% CI for $\hat{y}$ is

$$\hat{y} \pm t_{\alpha/2,n-(k+1)} \cdot s_{\hat{y}} = 42.253 \pm 2.06(.350) = (41.532, 42.974).$$

A 95% PI for a future y value is

$$\hat{y} \pm t_{\alpha/2,n-(k+1)} \cdot \sqrt{s^2 + s_{\hat{y}}^2} = 42.253 \pm 2.06\sqrt{1.09412^2 + 0.34^2} = (39.887, 44.619).$$

(f) A 99% CI for β_3 is

$$\hat{\beta}_3 \pm t_{\alpha/2,n-(k+1)} \cdot s_{\hat{\beta}_3} = -.04304 \pm 2.797(.01773) = (-.09263, .00655)$$

Since the CI includes zero, the variable should not be retained in the model. (Alternatively, the t-ratio is 2.4275, which compared to the critical value of 2.797 is not significant, hence failing to reject the null hypothesis of $\beta_3 = 0$.)