STAT/MATH 3340
REGRESSION ANALYSIS
(Fall 2006)

Lecture: MWF 930-1030 AM
in LSC 4258 (Psychology Wing)

Instructor: Michael Dowd
Office: Chase 116
Phone: 494-1048
e-mail: mdowd@mathstat.dal.ca
Office Hours: MW 1:00-3:00pm

Course Website:

`www.mathstat.dal.ca/~mdowd/stat3340/stat3340.html`

Copies of assignments, computer code, data sets and (selected) lecture notes will be made available on course website. Announcements will also be posted here.

University Policies

Students with Disabilities:

Students with disabilities should register as quickly as possible at Student Accessibility Services if they want to receive academic (classroom/exam) accomodations. To do so please phone 494-2836, email *disabilities@dal.ca*, or drop in at the Killam G28.

`http://studentaccessibility.dal.ca/`

Intellectual Honesty

`http://plagiarism.dal.ca/policies/index.html`

Method of Evaluation:

- Assignments 20%
- Project 20%
- Midterm Exam 15%
- Final Exam 45 %

(Note: no supplemental examination)

Grading Scheme:

A+	93 and above
A	85 up to but not including 93
A-	80 up to but not including 85
B+	75 up to but not including 80
B	70 up to but not including 75
B-	65 up to but not including 70
C+	60 up to but not including 65
C	55 up to but not including 60
C-	50 up to but not including 55
D	47 up to but not including 50
F	below 47

Formal Prerequisites:

1. STAT 2080 - Statistical methods for data analysis and inference

2. MATH 2030 - Linear algebra

3. MATH 1010 (Calculus) or STAT 2060 (Introduction to Prob and Stats)

Skills Required:

- Basic concepts from introductory statistics (probability, random variables, expectations and inference)

- A working knowledge of matrix algebra and basic knowledge of calculus.

Course Text: This is a required text and will be followed closely.

Introduction to Linear Regression Analysis
by Montgomery, Peck and Vining
(Third edition ... or more recent)

(Available in Dal Bookstore)

Statistical Software:

Assignments and projects will require some use statistical software (S-Plus) to carry out regression analyses

Course Goals:

Regression is a widely used statistical technique for investigating the relationship between variables. This course provides a thorough treatment of the theory and practice of regression analysis.

Objectives:

1. Develop an understanding of the theoretical basis for regression analysis

2. Develop the ability to (properly) carry out practical applications of regression analysis

3. Gain familarity with use of modern statistical software packages.

Topic Outline

- Simple linear regression
- Multiple linear regression and its matrix formulation
- Assessing model adequacy and transformations
- Regression diagnostics (influential points and multicollinearity)
- Special types of regression: polynomial and indicator variable regression
- Model building and variable selection
- Special topics (generalized linear models, nonlinear regression) - time permitting

Course Evalulation Details

Assignments:

There will be regular assignments worth 20% of the total mark. Late assignments will be given a zero mark unless a medical excuse is presented. Assignments will require both analytic work, as well as the use of statistical software.

Project:

There will be a project worth 20% of the total mark. The project will require written report.

Exams:

The midterm will take place during regularly scheduled class time. The final will take place during the Dalhousie exam period as scheduled by the registrar. Further details for these exams will be announced in class

DALHOUSIE UNIVERSITY FACULTY OF SCIENCE

Department of Mathematics and Statistics

MATH 3340 / STATISTICS 3340 : Final Examination

Date and Time: 1900-2200 pm, Dec 11, 2006

Name: _____ **Student ID:** _____

Signature: _____

This final exam is marked out of 100 points. The number of points allocated to each problem is indicated in the left margin. The time available is 3 hours. This is a closed book exam but you may refer to one 8.5x11 in. formula sheets (double sided), and you may use a calculator. Please check that there are 9 pages of questions, as well as t and F tables.

(14) 1. Consider the regression $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I})$. Given that

$$\mathbf{X} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

(a) Compute the least squares estimates $\hat{\boldsymbol{\beta}}$

(b) Determine $\hat{\mathbf{y}}$ and the residuals $\mathbf{e}$.

(c) Determine the variance-covariance matrix for $\hat{\boldsymbol{\beta}}$, i.e. $\text{cov}(\hat{\boldsymbol{\beta}})$. State the estimated variances for $\hat{\beta}_0$ and $\hat{\beta}_1$.

(13) 2. Given the multiple regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2 + \beta_3 x_1 x_2 + \varepsilon$. Suppose a regression was carried out using 13 observations on y. Suppose that $R^2 = 0.8$, and that $\sum_{i=1}^{n} e_i^2 = 48$.

(a) Fill out the regression ANOVA table below.

Source of variation	Sum of Squares	deg freedom	Mean Square	F_0
Regression				
Residual				–
Total			–	–

(b) Is the overall regression significant? Test $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ versus H_a : at least one of $\beta_j \neq 0$. Use a significance level of $\alpha = 0.01$. Show your work.

(c) The table below shows the results from the above regression for the estimated β_j for $j = 0, 1, 2, 3$. Carry out hypothesis tests on each of the individual regression coefficients (i.e. $H_0 : \beta_j = 0$ vs. $H_a : \beta_j \neq 0$) by filling in the values for the test statistics (t-value), and the associated P-value. Be sure to indicate which t distribution you are using for the hypothesis testing.

Coeff	Estimated Value	Standard Error	t value	P-value
β_0	71.65	14.14		
β_1	1.452	0.117		
β_2	0.416		2.24	
β_3	-0.237			0.1

(7) 3. Suppose that polynomial regression $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_{n-1} x^{n-1} + \varepsilon$ was fit to a data set where y is comprised of n observations. (Hint: consider this problem for the simple linear regression case when $n = 2$, i.e. fitting a straight line to two data points).

(a) What is R^2? Why?

(b) What is the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$? Why?

(8) 4. Contrast matrix regression for (i) ordinary least squares regression (the usual case), (ii) weighted least squares, and (iii) generalized least squares. For what situations would you use weighted least squares? For what case would you use generalized least squares?

(10) 5. Given the scaled and centered $\mathbf{X}'\mathbf{X}$ matrix (the correlation matrix):

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & \alpha \\ \alpha & 1 \end{bmatrix}$$

(a) Compute the expression for the variance inflation factors, C_{jj}, where $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}$. How large a value for α is required for the regression to have significant multicollinearity (as defined by a $C_{jj} > 10$)?

(b) Determine the expression for the condition number of $\mathbf{X}'\mathbf{X}$. How large a value for α would be required for the matrix to have significant multicollinearity (as defined by a condition number greater than 1000)?

(11) 6. Consider the multiple regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

with the usual assumptions: $E(\varepsilon) = 0$, $\text{var}(\varepsilon) = \sigma^2$, and ε uncorrelated. The full regression model was fit to 13 observations yielding $SS_{Res} = 47.9$. The following extra sums of squares were also obtained.

$$\begin{aligned} SS_R(\beta_1|\beta_0) &= 1809.4 \\ SS_R(\beta_2|\beta_1, \beta_0) &= 490.9 \\ SS_R(\beta_3|\beta_2, \beta_1, \beta_0) &= 341.6 \\ SS_R(\beta_4|\beta_3, \beta_2, \beta_1, \beta_0) &= 26.0 \end{aligned}$$

(a) Compute the sum of squares for the regression for the full model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$.

(b) What is $SS_R(\beta_1, \beta_0)$?

(c) What is $SS_R(\beta_4, \beta_3|\beta_2, \beta_1, \beta_0)$?

(d) Using the appropriate extra sum of squares, test the null hypothesis $H_0 : \beta_4 = 0$ versus the alternative $H_a : \beta_4 \neq 0$. Use a significance level of $\alpha = 0.05$.

(8) 7. Consider generalized least squares regression $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{V})$. Show that the variance-covariance matrix for the $\hat{\boldsymbol{\beta}}$ is $\sigma^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$, starting with the definition $\text{var}(\hat{\boldsymbol{\beta}}) = E\{(\hat{\boldsymbol{\beta}} - E\{\hat{\boldsymbol{\beta}}\})(\hat{\boldsymbol{\beta}} - E\{\hat{\boldsymbol{\beta}}\})'\}$. (Recall that $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$ and that $E\{\hat{\boldsymbol{\beta}}\} = \boldsymbol{\beta}$).

(8) 8. Consider the second order polynomial regression model $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$. Derive the least squares normal equations used to find the minimum of $S(\beta_0, \beta_1, \beta_2) = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i - \beta_2 x_i^2)^2$. (Note that only have to derive the system of equations, you do not have to solve it.)

(6) 9. Show that the least-squares estimator for ridge regression $\hat{\beta}_R = (\mathbf{X}'\mathbf{X} + \gamma\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}$ is a biased estimator (assume $\gamma > 0$).

(7) 10. Consider the matrix regression $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $E(\boldsymbol{\varepsilon}) = 0$ and $\text{var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$. Define $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$. Show the following:

(a) $\mathbf{e} - (\mathbf{I} - \mathbf{H})\mathbf{y} = \mathbf{0}$.

(b) $\mathbf{e}'\mathbf{e} = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}$.

(8) 11. Discuss how you would go about detecting influential observations in a regression analysis. Compare and contrast the methods discussed in the course.