MATH 423 Final Examination December 9th, 2008

Student Name:

Student Number:

McGill University
Faculty of Science
FINAL EXAMINATION

MATH 423
Regression and Analysis of Variance
December 9th, 2008
9 a.m. - 12 Noon

Calculators are allowed.

One 8.5" $\times$ 11" two-sided sheet of notes is allowed.

All language dictionaries are allowed.

Answer all your questions in the exam booklet provided.
You must do BOTH of the first two questions (Questions 1 and 2) and TWO of the remaining THREE questions (Questions 3 through 5) . You will receive points from the two questions with the highest marks from Questions 3, 4, and 5.

Question 6 is worth 10 bonus points. You will not lose points on the exam for any work on this question, you can only add points for any correct work that you provide.

There are 11 pages to this exam. The total number of marks for the exam is 100, although it is possible score as high as 110 due to the bonus question.

Examiner: Professor Russell Steele
Associate Examiner: Professor David Stephens

MATH 423 Final Examination December 9th, 2008

Question 1: (30 points)

The data for this analysis concern salary and other characteristics of all faculty in a small Midwestern college collected in the early 1980s for presentation in legal proceedings for which discrimination against women in salary was at issue. All persons in the data hold tenured or tenure track positions; temporary faculty are not included. The data were collected from personnel files and consist of the following quantities:

- `Sex:` 1 for female and 0 for male
- `Rank:`, 1 for Assistant Professor, 2 for Associate Professor, and 3 for Full Professor
- `Year:` Number of years in current rank
- `Salary:` Academic year salary in dollars

(a) Test for significance of gender, without controlling for the other variables. Refer clearly to the part(s) of the output that you are using for your tests. Use a significance level of $\alpha = 0.10$ for the test.

(b) Test for a significant of gender after controlling for rank and number of years in current rank. Use a significance level of $\alpha = 0.10$ for the test. Is your answer the same as in part (a)? Explain why or why not.

(c) Interpret the coefficients for `rank` in the model in part (b).

(d) Assess the validity of the model assumptions and the potential for misleading results due to influential points for model (2).

(e) Test whether the association of gender with the response depends on the rank.

Question 2: (20 points)

Assume that $y_i \sim \text{Binomial}(p_i, m)$ where $p_i = \beta_0 + x_i\beta_1$ for $i = 1, .., n$.

(a) List the assumptions of the usual linear regression model and whether or not they are satisfied for this situation.

(b) Prove that the usual simple linear regression estimator $\hat{\beta}_1$ be unbiased for the true β_1.

(c) Find an expression for the variance of $\hat{\beta}_1$ using ordinary least squares for this problem.

(d) What are the fitted values from this regression estimates of? Will they potentially be difficult to interpret? Why or why not?

Question 3: (25 points)

Assume that we are testing between two multiple linear regression model,

$$\begin{aligned} \text{Model A:} \quad & \mathbf{y} = \mathbf{X_1}\beta_\mathbf{1} + \epsilon \\ \text{Model B:} \quad & \mathbf{y} = \mathbf{X_1}\beta_\mathbf{1} + \mathbf{X_2}\beta_\mathbf{2} + \epsilon \end{aligned}$$

where $\mathbf{y}$ and ϵ are $(n \times 1)$, the ϵ_i are independent and identically distributed and $\epsilon_i \sim$ N(0,σ^2). Also assume that $\mathbf{X_1}$ is $(n \times p)$ and $\mathbf{X_2}$ is $(n \times q)$.

(a) Write down the F-statistic that you would use to test the null hypothesis that $\beta_2 = \mathbf{0}$.

(b) Write down the the Mallow's C_p criterion values for Model A and Model B.

(c) Show that one of the two Mallow's C_p criterion values can be expressed in terms of the F-statistic for comparing the models that you wrote down in part (a).

(d) Discuss the implications of choosing a model according to the F-test as compared to Mallow's C_p based on your answer to (c).

Question 4: (25 points)

Assume that the model for the data is the standard multiple linear regression model,

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

where the ϵ_i are independent and identically distributed and $\epsilon_i \sim$ N(0,σ^2).

(a) Describe what EXACTLY is plotted in a partial regression (or added-variable) plot for a column of $\mathbf{X}$.

(b) Describe what EXACTLY is plotted in a partial residual (or component-residual) plot for a column of $\mathbf{X}$.

(c) When will the partial regression plot look exactly the same as the partial residual plot?

(d) Describe the difference between the notions of leverage and influence in a regression model and how one might measure each of these things.

Question 5: (25 points)
Assume that the model for the data is the standard multiple linear regression model,

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

where the ϵ_i are independent and identically distributed and $\epsilon_i \sim$ N(0,σ^2). Let H be the hat matrix for a regression of $\mathbf{y}$ on $\mathbf{X}$. Let U be a $n \times 1$ vector with 1 as its first element and 0's elsewhere.

(a) Consider using U (not y) as the response variable in a regression with $n \times p$ design matrix $\mathbf{X}$. Show that elements of the vector of fitted values from the regression of U on X are the h_{1j} $(j = 1, ..., n)$ elements of the usual hat matrix for design matrix X.

(b) Show the vector of residuals from the regression have $(1 - h_{ij})$ have $1 - h_{ij}$ as the first element and the other elements are $-h_{ij}$.

(c) Two matrices A and B are considered to be *orthogonal* if $AB = BA = 0$. Show that $I - H$ and H are orthogonal.

(d) Use the result in part (c) to show that as long a column for an intercept is included in $\mathbf{X}$, then the true slope of the regression of $\mathbf{e} = (\mathbf{y} - \hat{\mathbf{y}})$ on $\hat{\mathbf{y}}$ is 0, where $\hat{y}$ is the vector of fitted values from a regression of $\mathbf{y}$ on $\mathbf{X}$.

(e) What is the slope of the regression of e on y?

BONUS: Question 6 (up to 10 extra marks)

Assume that $y_i \sim$ Normal$(\mu_i, \sigma^2\mu_i^4)$ where $\mu_i = \beta_0 + x_i\beta_1$. Find the variance stabilizing transformation for y_i.

MATH 423 Final Examination December 9th, 2008

```
> ### Regression output for Question 1
>
> #### Model (1)
>
> model1<-lm(Salary~Sex,data=salary)
> summary(model1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    24697        938  26.330   <2e-16 ***
Sex1           -3340       1808  -1.847   0.0706 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 5782 on 50 degrees of freedom
Multiple R-Squared: 0.0639,Adjusted R-squared: 0.04518
F-statistic: 3.413 on 1 and 50 DF,  p-value: 0.0706

> anova(model1)
Analysis of Variance Table

Response: Salary
          Df     Sum Sq    Mean Sq F value Pr(>F)
Sex        1  114106220  114106220   3.413 0.0706 .
Residuals 50 1671623638   33432473
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
>
>
```

```
> ### Model (2)
>
> model2<-lm(Salary~Sex+Rank+Year,data=salary)
> summary(model2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 15906.81     797.49  19.946  < 2e-16 ***
Sex1          524.15     834.69   0.628    0.533
Rank2        4373.92     906.12   4.827 1.51e-05 ***
Rank3        9483.84     912.79  10.390 9.19e-14 ***
Year          390.94      75.38   5.186 4.47e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2418 on 47 degrees of freedom
Multiple R-Squared: 0.8462,Adjusted R-squared: 0.8331
F-statistic: 64.64 on 4 and 47 DF,  p-value: < 2.2e-16

> anova(model2)
Analysis of Variance Table

Response: Salary
          Df     Sum Sq    Mean Sq F value    Pr(>F)
Sex        1  114106220  114106220  19.524 5.819e-05 ***
Rank       2 1239752324  619876162 106.063 < 2.2e-16 ***
Year       1  157183229  157183229  26.895 4.473e-06 ***
Residuals 47  274688086    5844427
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
>
>
```

```
> ### Model (3)
>
> model3<-lm(Salary~Rank+Year,data=salary)
> summary(model3)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16203.27     638.68  25.370  < 2e-16 ***
Rank2        4262.28     882.89   4.828 1.45e-05 ***
Rank3        9454.52     905.83  10.437 6.12e-14 ***
Year          375.70      70.92   5.298 2.90e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2402 on 48 degrees of freedom
Multiple R-Squared: 0.8449,Adjusted R-squared: 0.8352
F-statistic: 87.15 on 3 and 48 DF,  p-value: < 2.2e-16

> anova(model3)
Analysis of Variance Table

Response: Salary
          Df     Sum Sq    Mean Sq F value    Pr(>F)
Rank       2 1346783800  673391900 116.692 < 2.2e-16 ***
Year       1  161953324  161953324  28.065 2.905e-06 ***
Residuals 48  276992734    5770682
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
>
>
```

```
> ## Model (4)
>
> model4<-lm(Salary~Rank*Sex + Year,data=salary)
> summary(model4)


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 15952.10     855.91  18.638  < 2e-16 ***
Rank2        4383.11    1063.99   4.119 0.000161 ***
Rank3        8975.97    1133.16   7.921 4.49e-10 ***
Sex1          244.50    1159.16   0.211 0.833894
Year          409.90      78.21   5.241 4.10e-06 ***
Rank2:Sex1  -1059.19    2188.78  -0.484 0.630791
Rank3:Sex1   1582.95    1836.99   0.862 0.393417
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2432 on 45 degrees of freedom
Multiple R-Squared: 0.8509,Adjusted R-squared: 0.831
F-statistic:  42.8 on 6 and 45 DF,  p-value: < 2.2e-16

> anova(model4)
Analysis of Variance Table

Response: Salary
          Df     Sum Sq    Mean Sq  F value    Pr(>F)
Rank       2 1346783800  673391900 113.8150 < 2.2e-16 ***
Sex        1    7074743    7074743   1.1958    0.2800
Year       1  157183229  157183229  26.5667 5.494e-06 ***
Rank:Sex   2    8443427    4221713   0.7135    0.4954
Residuals 45  266244659    5916548
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
> ### Comparisons
> anova(model1,model2)
Analysis of Variance Table

Model 1: Salary ~ Sex
Model 2: Salary ~ Sex + Rank + Year
  Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
1     50 1671623638
2     47  274688086  3 1396935552 79.673 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> anova(model3,model2)
Analysis of Variance Table

Model 1: Salary ~ Rank + Year
Model 2: Salary ~ Sex + Rank + Year
  Res.Df       RSS Df Sum of Sq      F Pr(>F)
1     48 276992734
2     47 274688086  1   2304648 0.3943 0.5331

> anova(model1,model3)
Analysis of Variance Table

Model 1: Salary ~ Sex
Model 2: Salary ~ Rank + Year
  Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
1     50 1671623638
2     48  276992734  2 1394630904 120.84 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
> anova(model4,model1)
Analysis of Variance Table

Model 1: Salary ~ Rank * Sex + Year
Model 2: Salary ~ Sex
  Res.Df        RSS Df   Sum of Sq      F    Pr(>F)
1     45   266244659
2     50  1671623638 -5 -1405378979 47.507 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
> anova(model4,model2)
Analysis of Variance Table

Model 1: Salary ~ Rank * Sex + Year
Model 2: Salary ~ Sex + Rank + Year
  Res.Df       RSS Df Sum of Sq      F Pr(>F)
1     45 266244659
2     47 274688086 -2  -8443427 0.7135 0.4954
> anova(model4,model3)
Analysis of Variance Table

Model 1: Salary ~ Rank * Sex + Year
Model 2: Salary ~ Rank + Year
  Res.Df       RSS Df Sum of Sq      F Pr(>F)
1     45 266244659
2     48 276992734 -3 -10748075 0.6055 0.6148

> ## Influence measures
> summary(influence.measures(model2))
Potentially influential observations of
 lm(formula = Salary ~ Sex + Rank + Year, data = salary) :

   dfb.1_ dfb.Sex1 dfb.Rnk2 dfb.Rnk3 dfb.Year dffit   cov.r   cook.d hat
1  -0.13   0.05    -0.06    -0.05     0.27     0.31    1.41_*  0.02   0.24
7  -0.01  -0.09    -0.06    -0.15     0.12    -0.21    1.32_*  0.01   0.18
24 -0.61   1.28_*   0.35     0.95     0.03     1.76_*  0.16_*  0.42   0.12
```

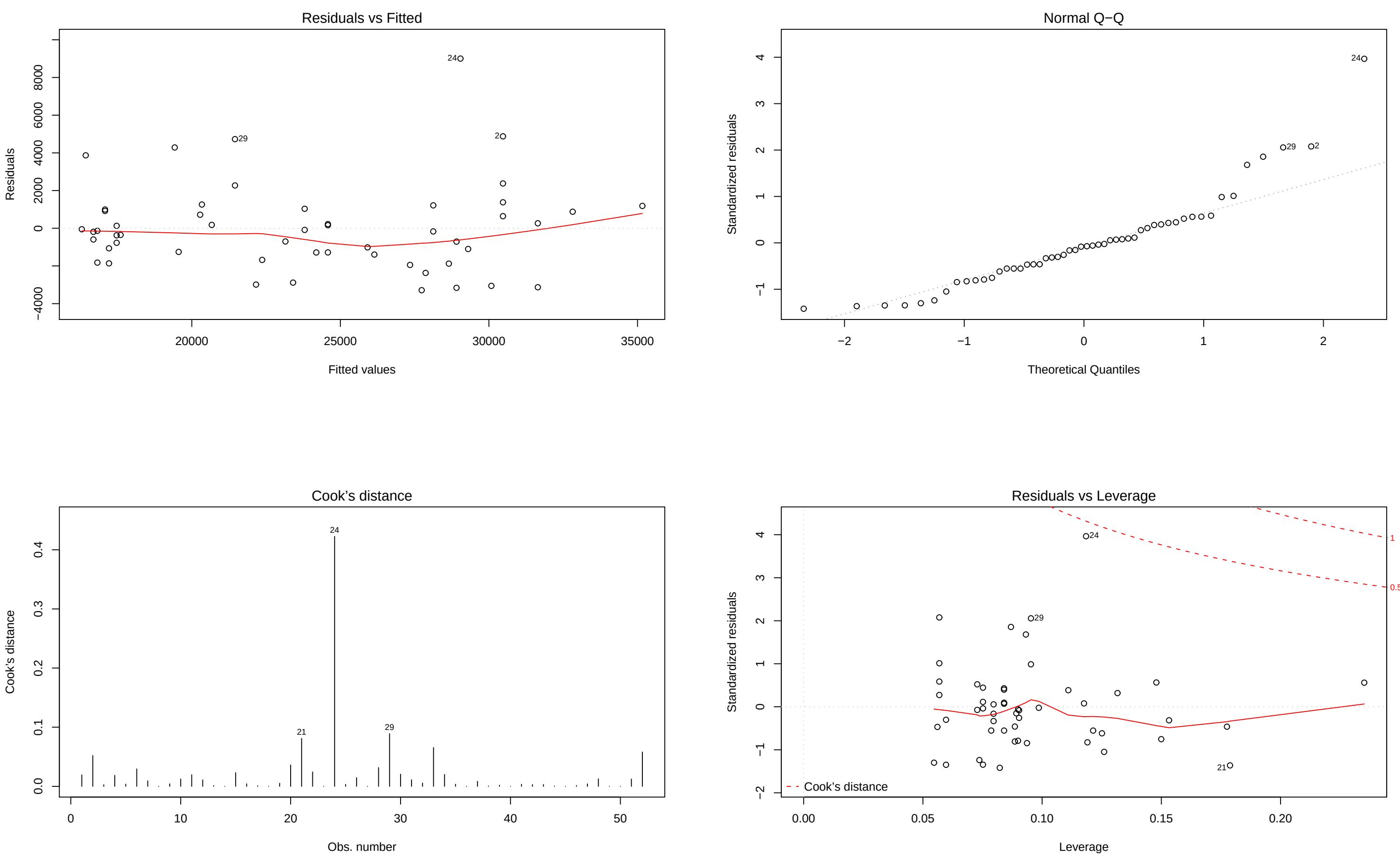

Figure 1: Regression diagnostics for Model 2