MATH 423 Final Examination December 11th, 2009

Student Name:

Student Number:

McGill University
Faculty of Science
FINAL EXAMINATION

MATH 423
Regression and Analysis of Variance
December 11th, 2008
9 a.m. - 12 Noon

Calculators are allowed.

One 8.5” × 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 11th, 2009

Question 1: (30 points)

Scleroderma is an auto-immune disease that causes inflammation of bodily tissue, including lung tissue. The activity of the disease is measured by a questionnaire that is scored between 0 and 10. Researchers are interested in modeling disease activity as a function of different aspects of the disease. The data set for the Take-Home Data Analysis contains information on **652** scleroderma patients in the Canadian Scleroderma Research Group registry. Assume that these patients are a random sample of the population of scleroderma patients.The regression output for all parts of this question is at the end of the exam.

- `LG33`: This is a measure of lung function, the Forced Vital Capacity (FVC), which is a percentage volume change of the lung between a full inspiration to total lung capacity.
- `ONRAYYR` - This is a measure of how long patients have had the disease (time since disease onset).
- `DEM2` - This is a gender variable.
- `DEM4` - This is a variable that gives the patient's age in years.
- `DEM69` - This is a variable that indicates a patient's smoking status
- `LAB28` - This is a test that indicates the absence or presence of Scleroderma-specific lung tissue damage detected during a chest X-ray. It has three possible values: "normal","abnormal", or "not done". The last case means that the patient did not receive a chest X-ray because the doctor did not think it was necessary. **Refer clearly to the part(s) of the output that you are using for your tests.**

(a) Test for significance of smoking without controlling for the other variables. Use a significance level of $\alpha = 0.01$ for the test.

(b) Explain the meaning of each of the coefficients for `model1`.

(c) Test for a significant of smoking after controlling only for age. Use a significance level of $\alpha = 0.01$ for the test. Is your answer the same as in part (a)? Explain why or why not.

(d) Interpret the coefficients for the status of the chest X-ray and smoking in `model3`.

(e) Which covariates in `model3` seem to be statistically significantly associated with the response? Assess the fit and the validity of the model assumptions for `model3`.

(f) Test whether the association of smoking with lung function depends on the status of the chest X-ray.

Question 2: (20 points)

Assume that we are fitting the simple linear regression model:

$$y_i = \beta_0 + x_i\beta_1 + \epsilon_i$$

where the ϵ_i are independent and identically distributed and $\epsilon_i \sim \mathrm{N}(0,\sigma^2)$ and $i = 1, ..., n$.

(a) Prove that the fitted simple linear regression line must pass through the point $(\bar{x}, \bar{y})$.

(b) Show that $\hat{y}_j = \sum_{i=1}^n h_{ij} y_i$, where $\hat{y}_j$ are the fitted values and

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

(c) Show that $\sum_{i=1}^n h_{ij} = 1$ and

$$\sum_{i=1}^n h_{ij} x_i = \bar{x} + \sum_{i=1}^n h_{ij}(x_i - \bar{x}).$$

(d) Using only the results in parts (b) and (c), show that $E(\hat{y}_j) = E(y_j)$.

Question 3: (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 $\epsilon_i \sim \mathrm{N}(0,\sigma^2)$.

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

(b) Show that the regression estimates constructed from the points in a partial residual plot will be the same whether an intercept is included or not included when calculating the slope from the partial residual plot.

(c) Compute the sample correlation between a column $\mathbf{x}_j$ and the partial residual vector for $\mathbf{x_j}$.

(d) What does your expression in part (c) suggest regarding the appearance of the partial residual plot when the R^2 value for the model with all covariates is very close to 1? How does this affect your interpretation of the partial residual plot?

MATH 423 Final Examination December 11th, 2009

Question 4: (25 points)

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

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

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

(a) Write down the forms of the Mallow's C_p criterion for Model A and Model B.

(b) Prove that the largest possible increase in C_p going from Model A to Model B is $2q$ for a fixed estimate of σ^2.

(c) Assume now that we have five available variables and that the best subset of size 3, in terms of Mallow's C_p, is a proper subset of the best subset of size 4. Determine a lower bound on the value of the F-statistic (with σ^2 estimated using all five variables) for testing that the coefficient of the added (fourth) regressor is zero if $C_3 - C_4 > 1$ and both C_p values are less than p.

(d) While recognizing that n and the significance level are not given, does your lower bound on the F-statistic suggest that a subset can be selected just based on the amount by which C_p is less than p? Why or why not?

MATH 423 Final Examination December 11th, 2009

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 $\epsilon_i \sim \mathrm{N}(0,\sigma^2)$.

(a) Find H where H is the matrix such that $Hy = \hat{y}$ where $\hat{y}$ are the fitted values from using ordinary least squares to estimate β above.

(b) Assume we plot the raw residuals vs. the fitted values for that particular model and observe significant heterogeneity in the variability of the residuals with respect to the fitted values. Using your answer to part (a), explain how the true model residuals, ϵ_i can have the same variance, but the raw residuals would still exhibit significant heterogeneity of variance.

(c) Suggest an alternate residual vs fitted plot that would solve the problem in part (b).

(d) Assume now that the only covariate of interest is an ordered categorical variable with three levels and that we use a set of orthogonal polynomial contrasts in our design matrix (e.g. two variables where the three levels of the categorical variable are encoded by (-1,0,1) and (1,-2,1) respectively). Assume as well that there are equal numbers of observations at each of the three levels of our covariate. Prove that the raw residual plot in part (b) will, in this case, be able to detect heterogeneity of variance in the residuals.

BONUS: Question 6 (up to 10 extra marks)

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 $\epsilon_i \sim \mathrm{N}(0,\sigma^2)$.

(a) Derive the steps of the Box-Tidwell procedure by using a first-order Taylor Series approximation of the mean of y in the model above.

(b) Explain the advantage(s) and disadvantages of using the Box-Tidwell procedure to determine the exponent of a regressor, rather than automatically using a linear term and a quadratic term.

```
> ### Regression output for Question 1
> model1<-lm(LG33~DEM69)
> summary(model1)

Call:
lm(formula = LG33 ~ DEM69)

Residuals:
    Min      1Q  Median      3Q     Max
-74.076 -14.076  -0.076  14.924  81.924

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                    74.076      1.393  53.185   <2e-16 ***
DEM69Smoked only in the past   -4.056      1.882  -2.156   0.0315 *
DEM69Current smoker            -4.096      2.615  -1.566   0.1177
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 22.02 on 649 degrees of freedom
Multiple R-squared: 0.008032,Adjusted R-squared: 0.004975
F-statistic: 2.627 on 2 and 649 DF,  p-value: 0.07303

> anova(model1)
Analysis of Variance Table

Response: LG33
           Df Sum Sq Mean Sq F value  Pr(>F)
DEM69       2   2548 1274.25  2.6275 0.07303 .
Residuals 649 314745  484.97
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
> model2a<-lm(LG33~DEM69+DEM4)
> model2b<-lm(LG33~DEM4)
> summary(model2)

Call:
lm(formula = LG33 ~ DEM69 + DEM4)

Residuals:
     Min       1Q   Median       3Q      Max
-72.3933 -14.5329   0.1067  14.0186  81.6730

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                  90.30181    4.10941  21.974  < 2e-16 ***
DEM69Smoked only in the past -3.96677    1.85818  -2.135   0.0332 *
DEM69Current smoker          -5.54417    2.60535  -2.128   0.0337 *
DEM4                         -0.29045    0.06932  -4.190 3.18e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 21.75 on 648 degrees of freedom
Multiple R-squared: 0.0342,Adjusted R-squared: 0.02973
F-statistic: 7.649 on 3 and 648 DF,  p-value: 4.972e-05

> anova(model2a)
Analysis of Variance Table

Response: LG33
           Df Sum Sq Mean Sq F value    Pr(>F)
DEM69       2   2548  1274.2  2.6945   0.06833 .
DEM4        1   8303  8302.8 17.5569 3.175e-05 ***
Residuals 648 306443   472.9
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
> anova(model1,model2a)
Analysis of Variance Table

Model 1: LG33 ~ DEM69
Model 2: LG33 ~ DEM69 + DEM4
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    649 314745
2    648 306443  1    8302.8 17.557 3.175e-05 ***
---
> anova(model2b,model2a)
Analysis of Variance Table

Model 1: LG33 ~ DEM4
Model 2: LG33 ~ DEM69 + DEM4
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    650 309535
2    648 306443  2    3092.9 3.2701 0.03863 *
---
> model3<-lm(LG33~DEM2+ONRAYYR+DEM4+LAB28+DEM69)
> summary(model3)

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   93.34379    3.91080  23.868  < 2e-16 ***
DEM2Male                      -4.26297    2.39036  -1.783  0.07499 .
ONRAYYR                       -0.26403    0.09188  -2.874  0.00419 **
DEM4                          -0.17809    0.06733  -2.645  0.00837 **
LAB28not normal              -16.45623    1.80164  -9.134  < 2e-16 ***
LAB28not done                 -4.79765    2.33637  -2.053  0.04043 *
DEM69Smoked only in the past  -3.14356    1.75252  -1.794  0.07332 .
DEM69Current smoker           -6.42197    2.46864  -2.601  0.00950 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 20.31 on 644 degrees of freedom
Multiple R-squared: 0.1628,Adjusted R-squared: 0.1537
F-statistic: 17.89 on 7 and 644 DF,  p-value: < 2.2e-16
```

```
> anova(model3)
Analysis of Variance Table

Response: LG33
           Df Sum Sq Mean Sq F value    Pr(>F)
DEM2        1   2661  2661.0  6.4513 0.0113210 *
ONRAYYR     1   6145  6145.3 14.8984 0.0001249 ***
DEM4        1   5800  5799.6 14.0602 0.0001930 ***
LAB28       2  33978 16989.1 41.1875 < 2.2e-16 ***
DEM69       2   3071  1535.5  3.7226 0.0246939 *
Residuals 644 265639   412.5
---
> model3b<-lm(LG33~DEM2+ONRAYYR+DEM4+DEM69+LAB28)
> anova(model3b)
Analysis of Variance Table

Response: LG33
           Df Sum Sq Mean Sq F value    Pr(>F)
DEM2        1   2661  2661.0  6.4513 0.0113210 *
ONRAYYR     1   6145  6145.3 14.8984 0.0001249 ***
DEM4        1   5800  5799.6 14.0602 0.0001930 ***
DEM69       2   2447  1223.5  2.9662 0.0522033 .
LAB28       2  34602 17301.1 41.9439 < 2.2e-16 ***
Residuals 644 265639   412.5
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
> model4<-lm(LG33~DEM69*LAB28+DEM2+ONRAYYR+DEM4+LAB28)
> summary(model4)

Coefficients:
                                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                                  92.35223    4.01041  23.028  < 2e-16 ***
DEM69Smoked only in the past                 -0.33785    2.43886  -0.139  0.88987
DEM69Current smoker                          -7.26411    3.23831  -2.243  0.02523 *
LAB28not normal                             -15.28223    2.86968  -5.325 1.40e-07 ***
LAB28not done                                 0.11881    3.82596   0.031  0.97524
DEM2Male                                     -4.05750    2.39332  -1.695  0.09050 .
ONRAYYR                                      -0.26914    0.09190  -2.929  0.00353 **
DEM4                                         -0.17911    0.06726  -2.663  0.00794 **
DEM69Smoked only in the past:LAB28not normal -3.95819    3.83822  -1.031  0.30281
DEM69Current smoker:LAB28not normal           5.13422    5.69409   0.902  0.36757
DEM69Smoked only in the past:LAB28not done   -9.70350    5.10699  -1.900  0.05788 .
DEM69Current smoker:LAB28not done            -2.73209    6.87699  -0.397  0.69129
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 20.28 on 640 degrees of freedom
Multiple R-squared: 0.1702,Adjusted R-squared: 0.156
F-statistic: 11.94 on 11 and 640 DF,  p-value: < 2.2e-16

> anova(model4)
Analysis of Variance Table

Response: LG33
             Df Sum Sq Mean Sq F value    Pr(>F)
DEM69         2   2548  1274.2  3.0976 0.0458362 *
LAB28         2  40242 20121.1 48.9126 < 2.2e-16 ***
DEM2          1    751   750.5  1.8244 0.1772624
ONRAYYR       1   5229  5228.6 12.7103 0.0003907 ***
DEM4          1   2885  2885.3  7.0138 0.0082878 **
DEM69:LAB28   4   2362   590.6  1.4356 0.2205987
Residuals   640 263276   411.4
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

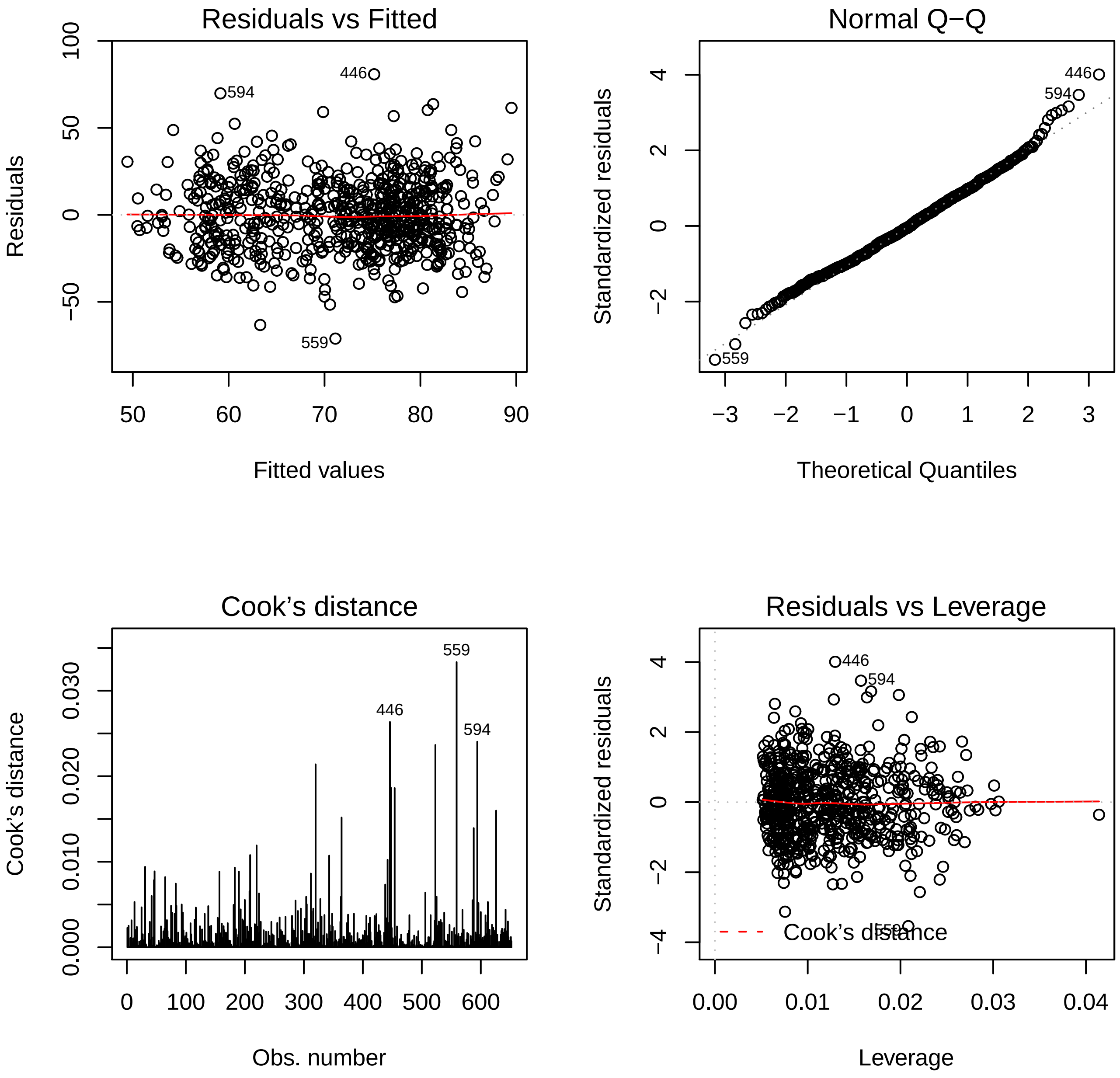

Figure 1: Regression diagnostics for `model3`