MATH 533 Final Examination December 11th, 2009

Student Name:

Student Number:

McGill University
Faculty of Science
FINAL EXAMINATION

MATH 533
Regression and Analysis of Variance
December 11th, 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 12 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 533 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.

MATH 533 Final Examination December 11th, 2009

Question 2: (20 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 $\epsilon_i \sim \text{N}(0,\sigma^2)$. Also assume that $\mathbf{X_1}$ is $(n \times p)$ and $\mathbf{X_2}$ is $(n \times q)$.

(a) Under the null hypothesis that Model A is the true model, derive the distribution of s_A^2, where s_A^2 is the residual standard error for model A.

(b) Show that we can write the difference in residual sums of squares for Model A as:

$$RSS_A = RSS_B + \hat{\beta}_2 \mathbf{X_2^t}(\mathbf{I} - \mathbf{H_1})\mathbf{X_2}\hat{\beta}_2$$

where $H_1 = \mathbf{X_1}(\mathbf{X_1^t X_1})^{-1}\mathbf{X_1}^t$.

(c) Use this result to find the expected values of s_A^2 under the hypothesis that Model B is the true model.

MATH 533 Final Examination December 11th, 2009

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) Consider a linear function of $\mathbf{d}^t\beta$ of β. Show that the change in the estimate $\mathbf{d}^t\hat{\beta}$ when the ith observation is deleted is:

$$\mathbf{d}^t\hat{\beta}_{-i} - \mathbf{d}^t\hat{\beta} = (\mathbf{C}^t\mathbf{d})_i e_i/(1 - h_i i)$$

where $\mathbf{C} = (\mathbf{X^t X})^{-\mathbf{1}})\mathbf{X^t}$, $e_i = (y_i - \hat{y}_i)$ and h_{ii} is the $i-$th diagonal element of the hat matrix.

(b) Show that

$$\mathbf{Y}_D - \mathbf{X}_D\hat{\beta}_{-D} = (\mathbf{I} - \mathbf{H_D})^{-1}[\mathbf{Y}_D - \mathbf{X}_D\hat{\beta}_{-D}]$$

where $\mathbf{Y}_D$ and $\mathbf{X}_D$ are the response vector and rows of the design matrix for a set of D observations, $\hat{\beta}_{-D}$ is the least squares estimate for a regression excluding those D observations and $\mathbf{H_D} = \mathbf{X_D}(\mathbf{X^t X})^{-1}\mathbf{X}_D^t$.

The following result may be helpful:

Let A and B be nonsingular $m \times m$ and $n \times n$ matrices respectively. Let U be be $m \times n$ and V be $n \times n$. Then

$$(\mathbf{A} + \mathbf{UBV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}\left(\mathbf{B}^{-1} + \mathbf{VA}^{-1}\mathbf{U}\right)^{-1}\mathbf{VA}^{-1}$$

MATH 533 Final Examination December 11th, 2009

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 $\epsilon_i \sim \mathrm{N}(0,\sigma^2)$ and $\mathbf{X}$ is $(n \times p)$. Consider the following extensions of the added variable plot.

(a) Suppose we fit the regression model above. Define $\mathbf{H_a}$ as the hat matrix for the regression model that uses only the first $p-2$ columns of $\mathbf{X}$ (and assume that the intercept column is the first column of $\mathbf{X}$). Let $\mathbf{z_y}$ be the residuals from this reduced regression model. Now use $\mathbf{z_y}$ as the response and the matrix $(\mathbf{I}-\mathbf{H_a})(\mathbf{x_{p-1}}, \mathbf{x_p})$ as the design matrix. In other words, using the two variables: $(\mathbf{I}-\mathbf{H_a})\mathbf{x_{p-1}}$ and $(\mathbf{I}-\mathbf{H_a})\mathbf{x_p}$ as two covariates in a regression model where $\mathbf{z_y}$ is the response. Show that the residuals from this regression are the same as the residuals that are obtained when fitting the full regression model with all p covariates.

(b) Under the same conditions as part (a), show that the regression coefficients for the regression of $\mathbf{z_y}$ on $(\mathbf{I}-\mathbf{H_a})\mathbf{x_{p-1}}$ and $(\mathbf{I}-\mathbf{H_a})\mathbf{x_p}$ are the same as the regression coefficients of $\mathbf{x_{p-1}}$) and $\mathbf{x_p}$ in the full model.

(c) Describe how one might use the results of parts (a) and (b) to construct a three-dimensional version of the added variable plot and how you could interpret it.

(d) As an alternative to parts (a) - (c), suppose that we regress $\mathbf{z_y}$ on $(\mathbf{I}-\mathbf{H_a})(\mathbf{x_{p-1}}\hat{\beta}_{p-1} + \mathbf{x_p}\hat{\beta}_p)$ where $\hat{\beta}_{p-1}$ and $\hat{\beta}_p$ are the least squares estimates from the full model. Find the slope of the fitted line and the residuals for this simple linear regression model and contrast it with those of the for the full model.

MATH 533 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 $\epsilon_i \sim \text{N}(0,\sigma^2)$.

(a) Show that the first step in forward selection is equivalent to selecting the variable that is most highly correlated with the response.

(b) Assume that there are $k < p$ variables in a candidate regression model with residual sums of squares RSS_k. Let RSS_{k+1} be the RSS for a candidate model that adds a single covariate. Show that the variable that increases the difference $RSS_k - RSS_{k+1}$ by the greatest amount is the one that has the largest partial correlation with the response given the variables already in the model. (The partial correlation of a covariate $\mathbf{x_j}$ with the response given a set of variables, C, is the sample correlation between the residuals from regressing x_j on the variables in C and the residuals from regressing y on the variables in C.) Interpret this result with respect to forward selection.

(c) Derive a similar result to part (b) when looking at deleting a variable from a candidate model with $k + 1$ covariates during backwards elimination.

BONUS: Question 6 (up to 10 extra marks)
Assume that the model for the data is the standard multiple linear regression model under a Box-Cox transformation:

$$\mathbf{z}(\lambda) = \mathbf{X}\beta + \epsilon$$

where the ϵ_i are independent and identically distributed and $\epsilon_i \sim \text{N}(0,\sigma^2)$ and

$$\mathbf{z}(\lambda) = \begin{cases} \frac{\mathbf{y}^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log(\mathbf{y}) & \lambda = 0 \end{cases}$$

(a) Perform a first order Taylor approximation of $z(\lambda)$ around $\lambda = 1$ and write it in terms of y and a function $u(y)$.

(b) Substituting your approximation of $z(\lambda)$ in part (a) into the regression equation, derive an approach for testing $\lambda = 0$ using least squares (this is known as Atkinson's method of choosing the power for the Box-Cox transformation).

MATH 533 Final Examination December 11th, 2009

```
> ### 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(model2a)

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
```

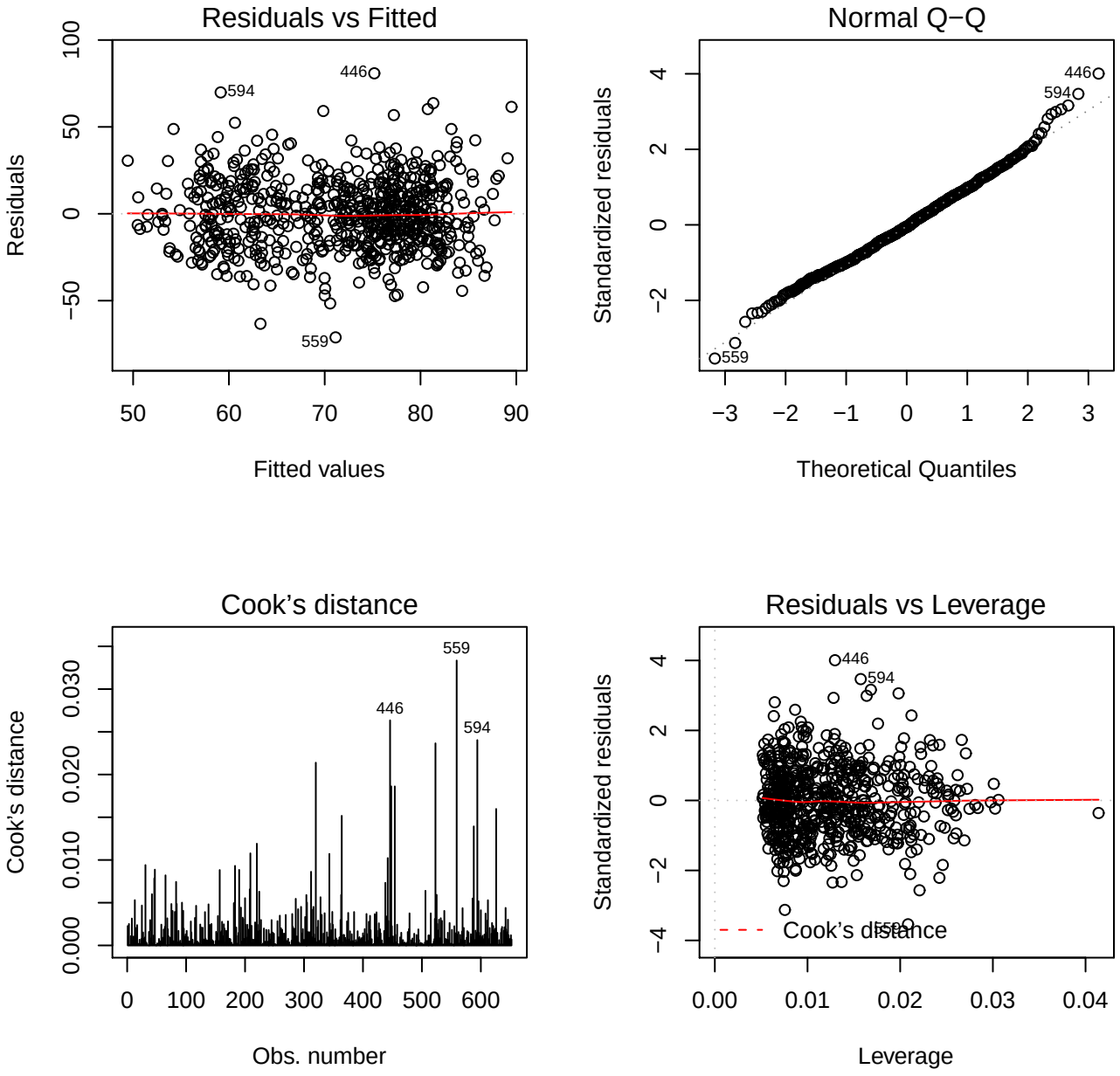

Figure 1: Regression diagnostics for `model3`