STATS 500 - Info and Practice Problems for Exam 2

December 2, 2015

Information: The Second Exam will be Thursday, December 10 from from 6:00-8:00pm in **Auditorium C in Angell Hall**. It will cover all the material in Chapters 7-11,13-15 of textbook **Linear Models in R** and Chapter 2 of textbook **Extending the Linear Model with R** as covered in the Lectures. The exam is open textbook (only the two textbooks) and open notes. Below is a set of practice Exam Problems. Other example problems to work on are the (undone) problems/exercises presented in class. Solutions to these practice problems will be posted by Monday, December 7. There will be a review session on Wednesday, December 9, in class.

Instructions: Attempt all questions. Show intermediate steps if there are any. Give only one answer per question.

Questions 1–4: These data are crime-related and demographic statistics for 47 U.S. states in 1960. The data are collected from the FBI's Uniform Crime Report and other government agencies to determine how the variable crime rate depends on the other variables measured in the study. The response variable is crime rate (`R`) – number of offenses reported to police per million population. There are 12 quantitative predictors describing the demographics of the state:

- `Age`: the number of males of age 14-24 per 1000 population)
- `Ed`: mean number of years of schooling ×10 for persons of age 25 or older
- `Ex0`: 1960 per capita expenditure on police by state and local government
- `Ex1`: 1959 per capita expenditure on police by state and local government
- `LF`: Labor force participation rate per 1000 civilian urban males age $14-24$
- `M`: The number of males per 1000 females
- `N`: State population size in hundred thousands
- `NW`: The number of non-whites per 1000 population
- `U1`: Unemployment rate of urban males per 1000 of age $14-24$
- `U2`: Unemployment rate of urban males per 1000 of age $35-39$
- `W`: Median value of transferable goods and assets or family income in tens of dollars
- `X`: The number of families per 1000 earning below 1/2 the median income

1. The correlation matrix of the predictors is given below. Name two features of this dataset that would lead you to consider applying shrinkage methods. Be specific.

```
        Age      Ed     Ex0     Ex1     LF       M
Age  1.0000 -0.5300 -0.5060 -0.5130 -0.161 -0.0287
Ed  -0.5300  1.0000  0.4830  0.4990  0.561  0.4370
Ex0 -0.5060  0.4830  1.0000  0.9940  0.121  0.0338
Ex1 -0.5130  0.4990  0.9940  1.0000  0.106  0.0228
LF  -0.1610  0.5610  0.1210  0.1060  1.000  0.5140
M   -0.0287  0.4370  0.0338  0.0228  0.514  1.0000
N   -0.2810 -0.0172  0.5260  0.5140 -0.124 -0.4110
NW   0.5930 -0.6650 -0.2140 -0.2190 -0.341 -0.3270
U1  -0.2240  0.0181 -0.0437 -0.0517 -0.229  0.3520
U2  -0.2450 -0.2160  0.1850  0.1690 -0.421 -0.0187
W   -0.6700  0.7360  0.7870  0.7940  0.295  0.1800
X    0.6390 -0.7690 -0.6310 -0.6480 -0.270 -0.1670

          N      NW      U1      U2       W       X
Age -0.2810  0.5930 -0.2240 -0.2450 -0.6700  0.6390
Ed  -0.0172 -0.6650  0.0181 -0.2160  0.7360 -0.7690
Ex0  0.5260 -0.2140 -0.0437  0.1850  0.7870 -0.6310
Ex1  0.5140 -0.2190 -0.0517  0.1690  0.7940 -0.6480
LF  -0.1240 -0.3410 -0.2290 -0.4210  0.2950 -0.2700
M   -0.4110 -0.3270  0.3520 -0.0187  0.1800 -0.1670
N    1.0000  0.0952 -0.0381  0.2700  0.3080 -0.1260
NW   0.0952  1.0000 -0.1560  0.0809 -0.5900  0.6770
U1  -0.0381 -0.1560  1.0000  0.7460  0.0449 -0.0638
U2   0.2700  0.0809  0.7460  1.0000  0.0921  0.0157
W    0.3080 -0.5900  0.0449  0.0921  1.0000 -0.8840
X   -0.1260  0.6770 -0.0638  0.0157 -0.8840  1.0000
```

2. A C_p plot for the model with crime rate R as the response is shown in Figure 1(a), and the following output is obtained:

```
Selection Algorithm: exhaustive
         Age Ed  Ex0 Ex1 LF  M   N   NW  U1  U2  W   X  
1  ( 1 ) " " " " "*" " " " " " " " " " " " " " " " " " "
2  ( 1 ) " " " " "*" " " " " " " " " " " " " " " " " "*"
3  ( 1 ) " " "*" "*" " " " " " " " " " " " " " " " " "*"
4  ( 1 ) "*" "*" "*" " " " " " " " " " " " " " " " " "*"
5  ( 1 ) "*" "*" "*" " " " " " " " " " " " " "*" " " "*"
6  ( 1 ) "*" "*" "*" " " " " " " " " " " " " "*" "*" "*"
7  ( 1 ) "*" "*" "*" " " " " " " " " " " "*" "*" "*" "*"
8  ( 1 ) "*" "*" "*" " " " " "*" " " " " "*" "*" "*" "*"
```

Which model should we select on the basis of C_p?

3. Ridge regression was also applied and the plot is shown in Figure 1(b). What advantage do the GCV selected ridge regression estimates have over the least squares coefficients?

4. Principal components regression with cross-validation was applied, and the cross-validated RMSE plot is shown in Figure 1(c). (Note: the number of components 0 corresponds to a model with an intercept only). What number of components should we use based on cross-validation?

Questions 5–12 refer to the following situation. An article appeared in the journal *Nature* entitled *Sexual Activity and the Lifespan of Male Fruit Flies* by Linda Partridge and Marion Farquhar.

The sexual activity of the fruit flies was manipulated by supplying individual males with one or eight receptive virgin females per day. The longevity of these males was compared with that of two control types. The first control consisted of two sets of individual males kept with one or eight newly inseminated (pregnant) females. Newly inseminated females will not usually remate for at least two days, and thus served as a control for any effect of competition with the male for food or space. The second control was a set of individual males kept with no females. There were 25 males in each of the five groups, which were treated identically in the provision of fresh food.

The variables in the data were

- `longevity`: lifespan in days
- `thorax`: thorax (body) length in *mm*
- `treat`: a five level factor representing the treatment groups. The levels were labeled as follows: "00" – no females, "10" – one pregnant female, "80" – eight pregnant females, "11" – one virgin female, "81" – eight virgin females.

The purpose of study was to investigate differences among the treatments on lifespan after allowing for the effect of thorax length which is known to be positively correlated with lifespan. No interaction was found between thorax length and the treatment factor. A treatment coding was used to represent the treatment factor and the following model was fit

```
longevity ~ treat + thorax
```

with following output obtained:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -49.984     10.609  -4.711 6.73e-06
treat10        2.653      2.975   0.891   0.3745
treat11       -7.017      2.973  -2.361   0.0199
treat80        3.929      2.997   1.311   0.1923
treat81      -19.951      3.006  -6.636 1.00e-09
thorax       135.819     12.439  10.919  < 2e-16

Residual standard error: 10.51 on 119 degrees of freedom
Multiple R-Squared: 0.6564,   Adjusted R-squared: 0.6419
F-statistic: 45.46 on 5 and 119 DF,  p-value: 0
```

5. Assuming the same treatment, how much longer would you expect a fly with `thorax` length $0.1mm$ greater than another to live?

6. What is the predicted difference in `longevity` between a male fly kept with one pregnant female and one kept with eight virgin females assuming they have the same `thorax` length?

7. Explain how you would test whether the difference in the previous question was statistically significantly different from zero.

8. Because the flies were randomly assigned to the five groups, the distribution of `thorax` lengths in the five groups are essentially equal. What disadvantage would the investigators have incurred by ignoring the `thorax` length in their analysis (i.e. had they done a one-way ANOVA instead)?

9. The residual-fitted plot is shown in the left panel of Figure 2. What is the main problem indicated by this plot? What action could you take to address this problem?

10. Two cases had missing values of longevity and were deleted from the analysis. What assumption about the missing values is necessary for the conclusions to be valid?

11. The Box-Cox procedure was used to determine a good transformation for this data. The plot of log-likelihood for λ is shown in the right panel of Figure 2. What transformation should be used to improve the fit and yet retain some interpretability?

12. I refit the model with the transformation determined using the Box-Cox method above. The second question above asked: Assuming the same treatment, how much longer would you expect a fly with `thorax` length $0.1mm$ greater than another to live? If you knew what transformation I used, could you answer this question using this new model? Explain.

Questions 13–16 concern the following situation:

A plant scientist measured the concentration (`conc`) of a particular virus in plant sap. The scientist chose 13 potato clones (coded a through m) for the study. Plant sap was taken from 5 inoculated plants of each clone, for a total of 65 measurements. Unfortunately, one measurement was lost during processing the samples.

Here is the output of a linear model fit to the data. (Treatment contrasts were used).

```
> g <- lm(conc ~ code, potato)
> summary(g)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1358.2      106.8  12.715  < 2e-16
codeb        -1313.6      151.1  -8.696 1.21e-11
codec          431.1      160.2   2.690 0.009626
coded        -1093.4      151.1  -7.238 2.29e-09
codee         -930.2      151.1  -6.158 1.16e-07
codef         -712.8      151.1  -4.719 1.89e-05
codeg         -482.0      151.1  -3.191 0.002429
codeh        -1079.0      151.1  -7.143 3.24e-09
codei        -1070.4      151.1  -7.086 3.98e-09
codej         -611.8      151.1  -4.050 0.000175
codek         -952.0      151.1  -6.302 6.88e-08
codel         -290.0      151.1  -1.920 0.060488
codem        -1029.4      151.1  -6.815 1.07e-08

Residual standard error: 238.8 on 51 degrees of freedom
Multiple R-Squared: 0.8263,  Adjusted R-squared: 0.7855
F-statistic: 20.22 on 12 and 51 DF,  p-value: 2.558e-15
```

13. What is the estimated mean concentration for clone a?

14. State explicitly the null hypothesis corresponding to the p-value of 1.07e-08.

15. Which clone had the missing value?

16. A clone will be classified as susceptible if the estimated mean concentration exceeds 1000. Which clones should be classified as susceptible?

17. After seeing the results of the experiment, the scientist wishes to compare clones f and ℓ. Is there evidence of a difference between these two clones? ($q_{13,52}^{0.95} = 4.91$ and $t_{51}^{0.975} = 2.01$)

18. Based on the Bonferonni method, for which codes ”b-m”, can you say are different from code ”a” with an overall significance level of .01. Be clear to justify your answer. What can you say about what your answer would be using the FDR method?

Questions 19–22 concern the following situation:

Consider the Pima diabetes dataset where we fit a binomial regression model with the result of the diabetes test (test) as a response and pregnant, glucose, diastolic, bmi, diabetes and age as predictors. Below is a listing of the data for the first few women in the data set.

```
> pima
    pregnant glucose diastolic triceps insulin  bmi diabetes age test
1          6     148        72      35       0 33.6    0.627  50    1
2          1      85        66      29       0 26.6    0.351  31    0
3          8     183        64       0       0 23.3    0.672  32    1
4          1      89        66      23      94 28.1    0.167  21    0
5          0     137        40      35     168 43.1    2.288  33    1
:          :      :         :       :        :   :        :    :    :
```

Below is a summary of some of the predictor variables.

```
> summary(pima)
    insulin           bmi           diabetes           age
 Min.   :  0.0   Min.   : 0.00   Min.   :0.0780   Min.   :21.00
 1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437   1st Qu.:24.00
 Median : 30.5   Median :32.00   Median :0.3725   Median :29.00
 Mean   : 79.8   Mean   :31.99   Mean   :0.4719   Mean   :33.24
 3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262   3rd Qu.:41.00
 Max.   :846.0   Max.   :67.10   Max.   :2.4200   Max.   :81.00
```

Here is the output of this binomial regression model fit to the data.

```
> pimaglm = glm(cbind(test,1-test) ~ pregnant + glucose + diastolic + bmi +
diabetes + age,family = binomial, pima)

> summary(pimaglm)

Call:
glm(formula = cbind(test, 1 - test) ~ pregnant + glucose + diastolic +
    bmi + diabetes + age, family = binomial, data = pima)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7380  -0.7313  -0.4123   0.7276   2.8984

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.239812   0.701970 -11.738  < 2e-16 ***
pregnant     0.124919   0.031972   3.907 9.34e-05 ***
glucose      0.033492   0.003440   9.736  < 2e-16 ***
diastolic   -0.013487   0.005114  -2.637  0.00836 **
bmi          0.087676   0.014268   6.145 7.99e-10 ***
diabetes     0.896150   0.294862   3.039  0.00237 **
age          0.016325   0.009237   1.767  0.07719 .
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 993.48  on 767  degrees of freedom
Residual deviance: 725.46  on 761  degrees of freedom
AIC: 739.46
```

19. Can the deviance be used to test the goodness of fit here? Explain.

20. What is the difference in the odds of testing positive for a woman with a BMI at the first quartile compared to a woman with BMI at the third quartile, with all other predictors held constant?

21. Referring to Figure 3, so women who test positive for diabetes have higher diastolic blood pressure? Is the diastolic blood pressure significant in the regression model? Explain the difference between these two questions and why the answers only appear contradictory.

22. Predict the probability of testing positive for a 30-year old woman who has been pregnant once, has glucose measurement of 100, diastolic blood pressure 70, BMI 25, and diabetes pedigree measurement of 0.6.

Questions 23-24 are general and do not relate to the data above.

23. The dataset `discoveries` lists the number of ”great” inventions and scientific discoveries in each year from 1860 to 1959. How would you test whether the discovery rate remained constant over time? State explicitly our assumptions, proposed statistical analyses, and what

you would be looking for in the output from the analysis. Let $\{y_i\}_1^{100}$ denote the number of inventions in each of the years 1860, 1861,...., 1959.

24. Suppose have a regression equation give by

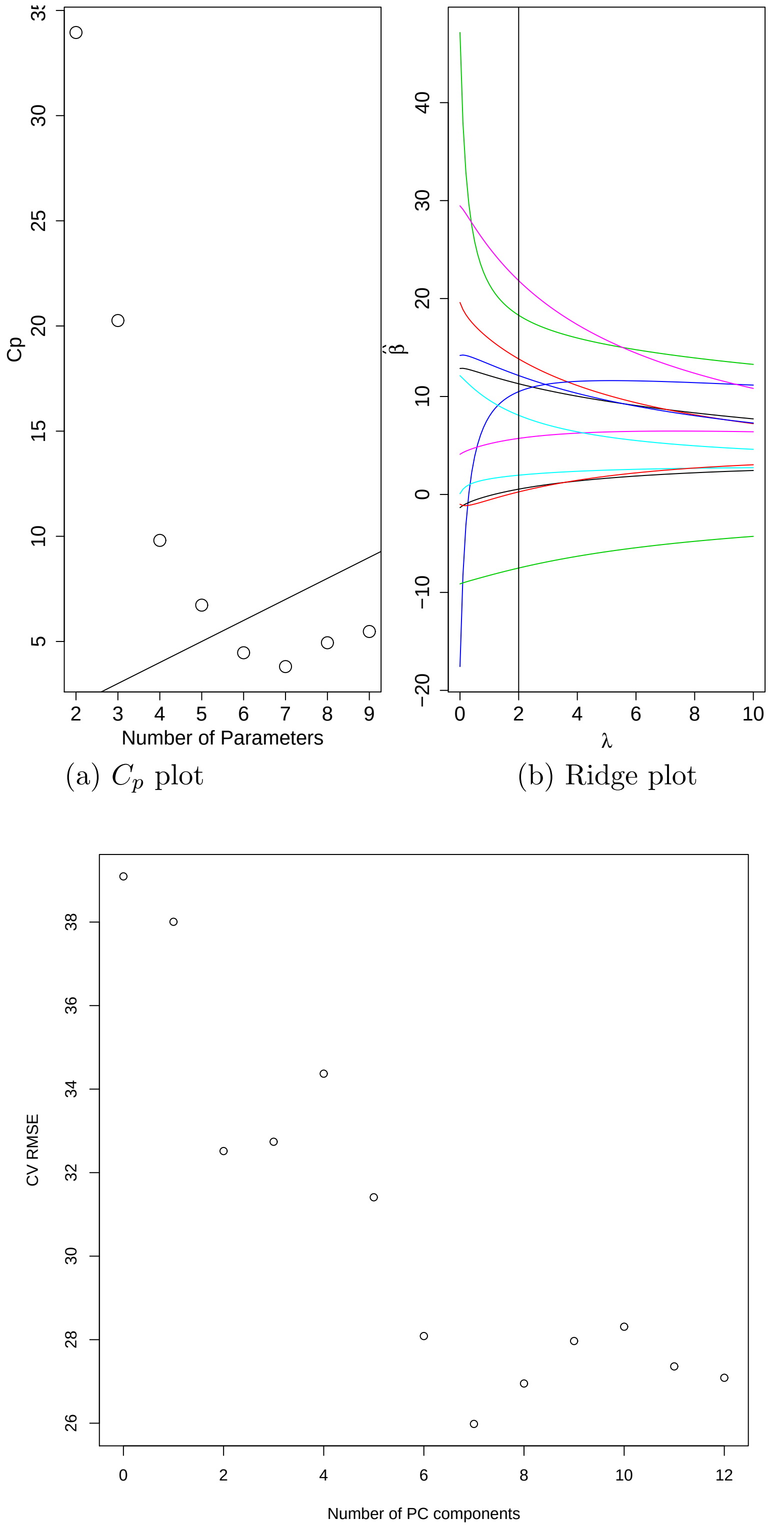
$$y_i = \beta_o + \beta_1 x_i + \epsilon_i \qquad\qquad i = 1, 2, \ldots, n$$

where the errors $\{\epsilon_i\}$ have mean 0, are independent, and $\text{Var}(\epsilon_i) = \kappa \cdot i$ for some constant $\kappa > 0$. Derive the form of the best linear unbiased estimator of

$$\beta = \begin{bmatrix} \beta_o \\ \beta_1 \end{bmatrix}$$

Give the covariance matrix of your estimator for β.

Hint: Give the answers in matrix form (with matrix products and inverses) and give explicit descriptions of all matrices.



(a) C_p plot

(b) Ridge plot

(c) PCR plot

Figure 1: Crime data plots

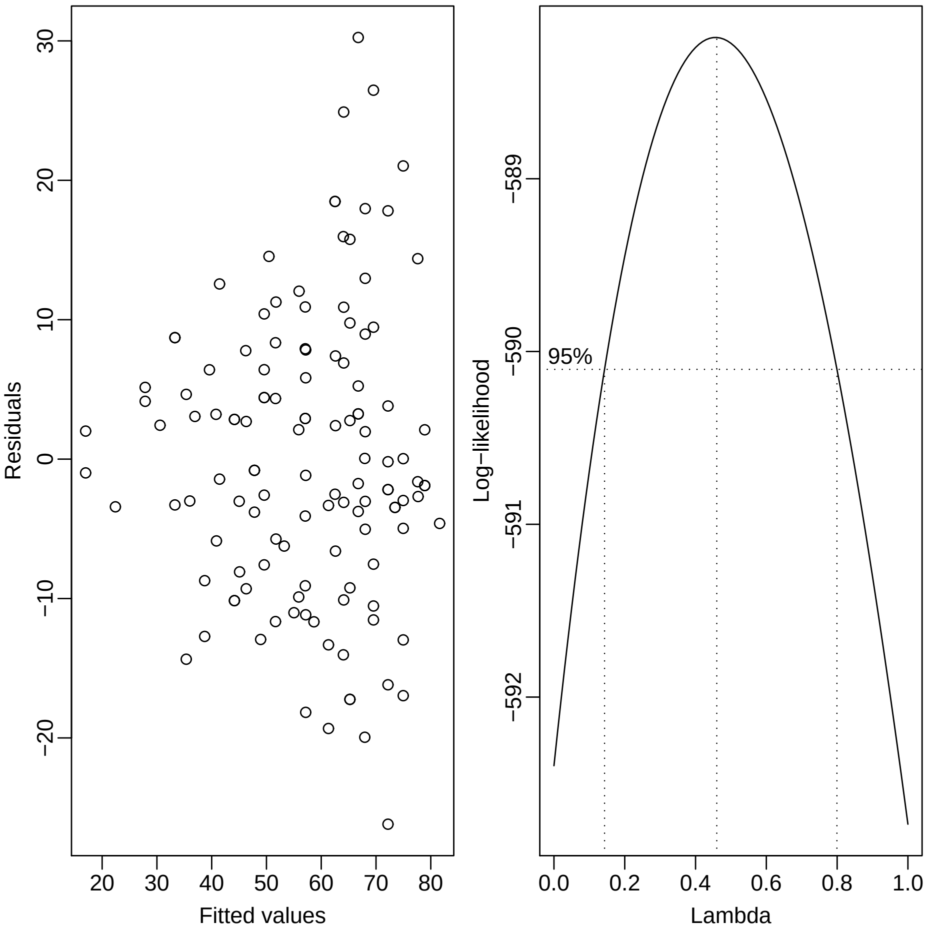


Figure 2: Longevity of Fruit Flies Plots

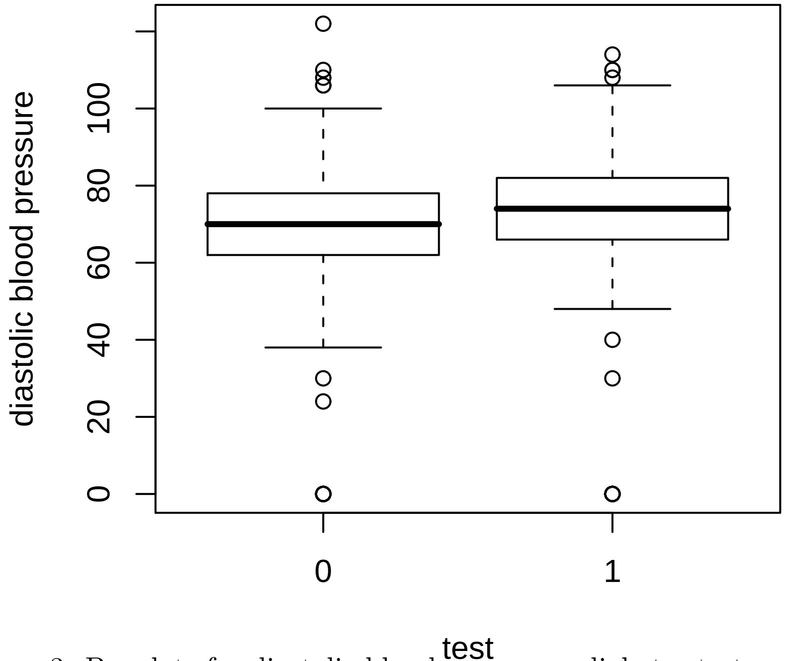


Figure 3: Boxplots for diastolic blood pressures: diabetes test = 0,1