MVE190/MSG500: Linear Statistical Models
Time: 08:30-12:30, Date: 2015-01-15

Instructor: José Sánchez (Rebecka Jörnsten)
Jour: José Sánchez, tel. 031–772 53 77.
Help: Course notes, your own notes, books.
Grading scale: Max points 35 (5 points each question).
Chalmers: 3 requires 14 points, 4 requires 21 points, 5 requires 28 points.
GU: G 14 points, VG 28 points.

Question 1

a) Let $\hat{\beta}_1$ and $\hat{\beta}_2$ be the least squares estimators of β from $y_1 = X\beta + \epsilon_1$ and $y_2 = X\beta + \epsilon_2$ respectively. If $\hat{\beta} = w\hat{\beta}_1 + (1-w)\hat{\beta}_2$, with $0 < w < 1$, determine the value of w that minimizes the variance of $\hat{\beta}$.

b) Discuss the statement: "In well-designed experiments with quantitative x-variables it is not necessary to use procedures for reducing the number of included x-variables after the data have been obtained".

Question 2

Consider a logistic regression model describing the relationship between students' math scores in high-school and the log odds of being member of a honors class. The results are shown in the table below.

```
Logistic regression                               Number of obs   =        200
                                                  LR chi2(1)      =      55.64
                                                  Prob > chi2     =     0.0000
Log likelihood = -83.536619                       Pseudo R2       =     0.2498

------------------------------------------------------------------------------
         hon |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        math |   .1563404   .0256095     6.10   0.000     .1061467     .206534
   intercept |  -9.793942   1.481745    -6.61   0.000    -12.69811   -6.889775
------------------------------------------------------------------------------
```

a) What are the odds of being in an honors class when the math score is zero?

b) How do we interpret the coefficient for "math"?

c) What is the expected increase in the odds of being in a honors class if the math score increases one unit?

Question 3

The data for this question comes from a study by Stamey et al. (1989). They examined the correlation between the level of prostate-specific antigen and certain clinical variables: log cancer volume (lcavol), log prostate weight (lweight), age, log of amount of benign prostatic hyperplasia (lbph), seminal vesicle invasion (svi) and percent of Gleason scores 4 or 5. The table below shows a summary of the results for different regression methods, namely ordinary least squares (LS), best subset selection of the OLS (Best Subset), Ridge regression (Ridge), lasso penalized regression (Lasso) and principal components regression (PCR).

Term	LS	Best Subset	Ridge	Lasso	PCR
Intercept	2.465	2.477	2.452	2.468	2.497
lcavol	0.680	0.740	0.420	0.533	0.543
lweight	0.263	0.316	0.238	0.169	0.289
age	−0.141		−0.046		−0.152
lbph	0.210		0.162	0.002	0.214
svi	0.305		0.227	0.094	0.315
lcp	−0.288		0.000		−0.051
gleason	−0.021		0.040		0.232
pgg45	0.267		0.133		−0.056
Test Error	0.521	0.492	0.492	0.479	0.449
Std Error	0.179	0.143	0.165	0.164	0.105

a) Discuss the main differences between these regression methods.

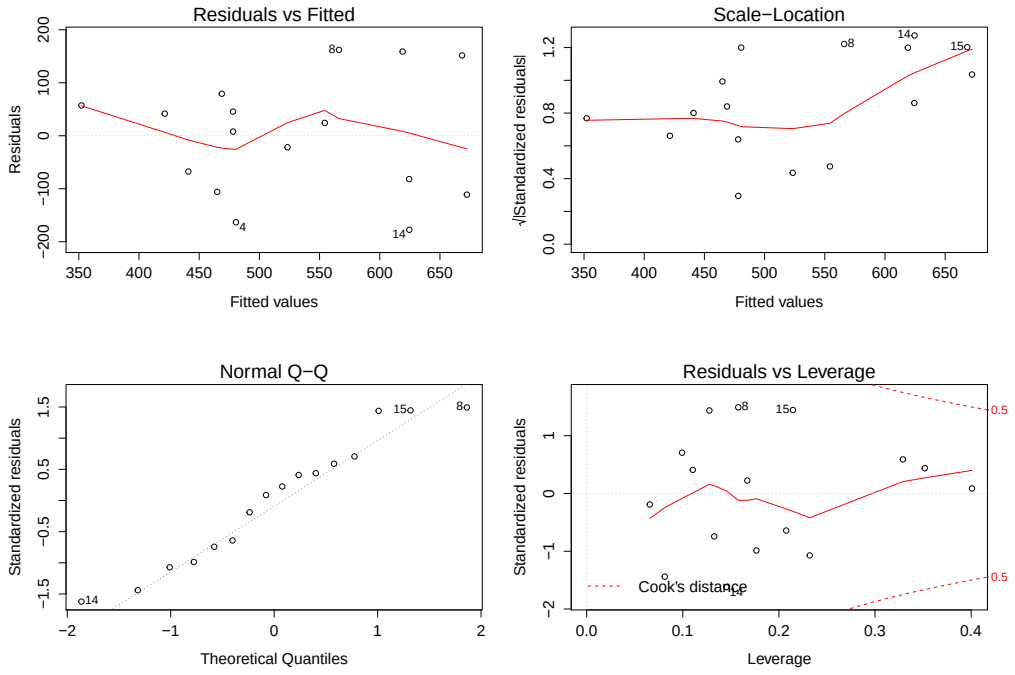
b) Discuss the results shown in the table.

Question 4

The table and figure below relate the a regression of the annual catch of Brevoortia patronus in the Gulf Menhaden measured in tons, to the number of vessels and fishing pressure for 1964 to 1979 (Nelson and Ahrenholz, 1986).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-119.0926	469.3404	-0.25	0.8037
vessels	2.2783	6.3870	0.36	0.7270
pressure	1.0538	0.3685	2.86	0.0134

Table 1: Regression summary

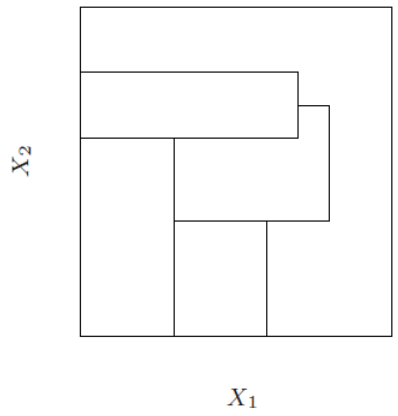


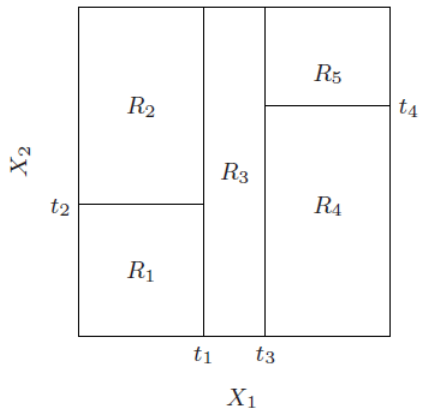
a) Comment on the diagnostic plots.

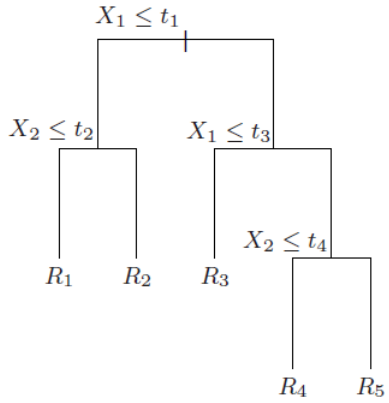
b) Comment on the regression results.

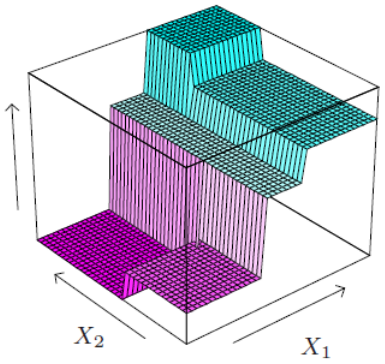
Question 5

Consider the figure below









a) Which of the figures at the top corresponds to a partition obtained by binary-splitting (as in CART)? Explain why.

b) The figures at the bottom correspond to a partition obtained by CART on simulated data. Explain how are they related to each other and how were they obtained.

Question 6

The following table shows the values for different model selection methods (AIC, BIC Mallow's Cp and R^2 adjusted for five models (M1 to M5).

Model	AIC	BIC	Cp	R^2_{adj}
M1	279	282	23	-
M2	266	270	4	0.37
M3	281	285	24	0.00
M4	268	273	6	0.35
M5	265	272	4	0.43

a) What is the difference between criteria like Mallow's Cp and R^2_{adj}?

b) Which model will each method choose in this example?

c) Can you suggest other variable selection methods?

d) How can you decide which model selection method to use?

Question 7

The following approach was used to determine the effect of acid rain on agricultural production. The U.S. Department of Agriculture statistics on crop production, fertilizer practices, insect control, fuel costs, land costs, equipment costs, labor costs, and so forth for each county in the geographical area of interest were paired with county-level estimates of average pH of rainfall for the year. A multiple regression analysis was run in which "production" (in dollars) was used as the dependent variable and all input costs plus pH of rainfall were used as independent variables. A stepwise regression analysis was used with pH forced to be in all regressions. The regression coefficient on pH from the model chosen by stepwise regression was taken as the measure of the impact of acid rain on crop production.

a) Discuss the validity of these data for establishing a causal relationship between acid rain and crop production.

b) Suppose a causal effect of acid rain on crop production had already been established from other research. Discuss the use of the regression coefficient for pH from these data to predict the change in crop production that would result if rain acidity were to be decreased. Do you see any reason the prediction might not be valid?

c) Suppose the regression coefficient for pH were significantly negative (higher pH predicts lower crop production). Do you see any problem with inferring that stricter government air pollution standards on industry would result in an increase in crop production?

d) Do you see any potential for bias in the estimate of the regression coefficient for pH resulting from the omission of other variables?