MVE190/MSG500: Linear Statistical Models
Time: 08:30-12:30, Date: 2014-04-25

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

Consider a data set (x_{i1}, x_{i2}, y_i), $i = 1, \ldots, n$, satisfying the equation:

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2}.$$

a) Derive the least squares estimate of α, β_1 and β_2 under the assumption that $\beta_1 = \beta_2$?

b) Do the same for the case $\beta_1 = 2\beta_2$

c) Is it possible to write down explicitly the solution for the case $\beta_1 > \beta_2$? Explain.

d) Suppose you compute the least squares estimates of β_1 and β_2 without any extra assumptions. How would you test the assumptions about β_1 and β_2 in a)-c) using the results from the standard least squares fit?

Question 2

Suppose you are studying the results of an experimental surgical procedure and you have data from 40 patients who underwent the procedure. The data consist of two variables: the age of the patient, x_i, and a categorical variable, y_i, that indicates if the patient died within 30 days of the procedure or not. Let θ_i be the probability that patient i survived.

a) Is it correct to model the probability of survival as $\theta_i = \beta_0 + \beta_1 x_i$? Explain.

b) Suppose you opt for the model $\theta_i = F(\beta_0 + \beta_1 x_i)$. What assumptions on F are required? Suggest an F function.

c) Suppose you choose to model θ_i as:

$$\theta_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

and that β_1=0.1. What does this result implies regarding the odds of death?

d) Write down the log-likelihood function for the logistic model.

Question 3

Tables 1, 2 and 3 show the the results' summary for three different models for the above data.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-10.4817	4.3041	-2.44	0.0149
age	0.1629	0.0702	2.32	0.0202

Table 1: Model 1

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.7334	0.3720	-1.97	0.0487
age	0.1629	0.0702	2.32	0.0202

Table 2: Model 2

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-142.6381	68.8146	-2.07	0.0382
age	4.4590	2.2069	2.02	0.0433
age2	-0.0347	0.0176	-1.97	0.0490

Table 3: Model 3

a) Comment about the fit of the models in general.

b) The difference between Model 1 and 2 is that the age has been centered. Comment on the effect of this.

c) In Model 3 a second predictor has been included, namely, the square of the age (age2). Comment on the effect of this.

d) When performing logistic regression, the estimates of β_0 and β_1 can be correlated. How would you check if this is the case for a certain model?

Question 4

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 (pgg45).

	lcavol	lweight	age	lbph	svi	lcp	gleason
lweght	0.300						
age	0.286	0.371					
lbph	0.063	0.437	0.287				
svi	0.593	0.181	0.129	-0.139			
lcp	0.692	0.157	0.173	-0.089	0.671		
gleason	0.426	0.024	0.366	0.033	0.307	0.476	
pgg45	0.483	0.074	0.276	-0.030	0.481	0.663	0.757

Table 4: Correlations of predictions for the prostate cancer data.

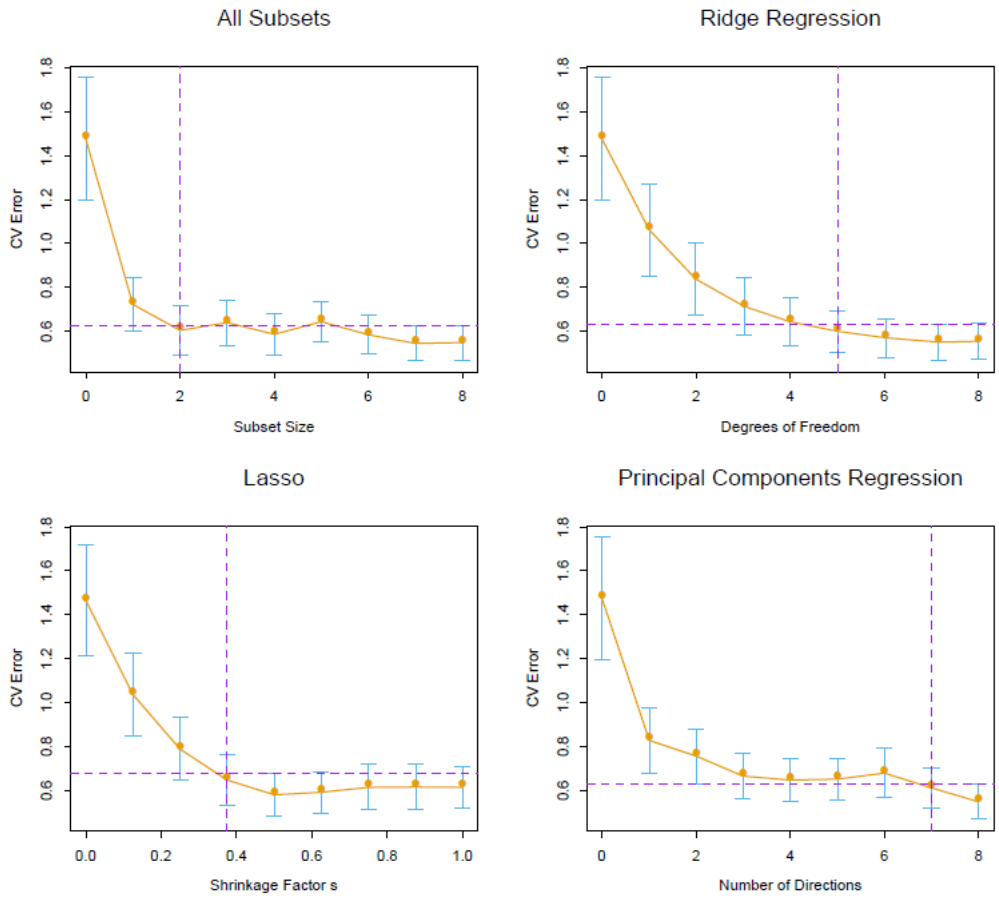

Figure 1: Results for model selection

Term	Coefficient	S.E.	Z score
Intercept	2.46	0.09	27.60
lcavol	0.68	0.13	5.37
lweight	0.26	0.10	2.75
age	-0.14	0.10	-1.40
lbph	0.21	0.10	2.06
svi	0.31	0.12	2.47
lcp	-0.29	0.15	-1.87
gleason	-0.02	0.15	-0.15
pgg45	0.27	0.15	1.74

Table 5: Linear model fit

a) Table 4 shows the correlations for the predictors and Table 5 shows a linear model fit. Comment on the results. What can you say about the significance of the estimates?

b) Figure 1 shows the results of four different methods for model selection. Comment briefly how each one of them is performed and interpret the results.

Question 5

Consider the multivariate regression model:

$$y = X\beta + \epsilon,$$

where $y, \epsilon \in \mathbb{R}^n$, $\beta \in \mathbb{R}^p$ and $X \in \mathbb{R}^{n \times p}$. Suppose that the ϵ_i are random with mean 0, constant variance σ^2 and that they are uncorrelated. The ordinary least squares estimator of β is given by $\hat{\beta} = \left(X^T X\right)^{-1} X^T y$. Let $\tilde{\beta} = \left[\left(X^T X\right)^{-1} X^T + D\right] y$ be another estimator of β, where D is a $n \times p$ nonzero matrix.

a) Derive the conditions that D has to satisfy in order for $\tilde{\beta}$ to be an unbiased estimate of β.

b) Suppose D fulfills the conditions to make $\tilde{\beta}$ an unbiased estimate of β. Show that the variance of $\tilde{\beta}$ is larger than the variance $\hat{\beta}$.

Question 6

In this question the strikes data from StatLib is used. The data consist of annual observations on the level of strike volume (days lost due to industrial disputes per 1000 wage salary earners), and their covariates in 18 OECD countries from 1951-1985. The 7 data fields include the following variables: country, year, strike volume (volume), unemployment, inflation, parliamentary representation of social democratic and labor parties (prep) and a time-invariant measure of union centralization (ucen).

a) Comment on the scatterplots and boxplots shown in Figures 2 and 3. If you want to predict strike volume based on the other variables, are there any transformations needed?

b) Table 6 shows the results of a linear model (observe that not all of the variables where included). Comment on the results.

c) Figure 4 shows the diagnostic plots. Comment on the results.

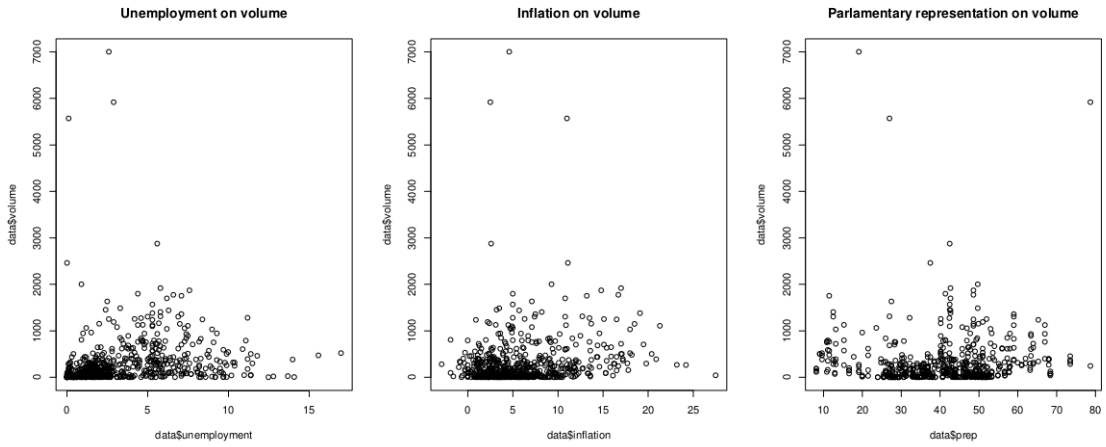

Figure 2: Scatterplots for continuous variables

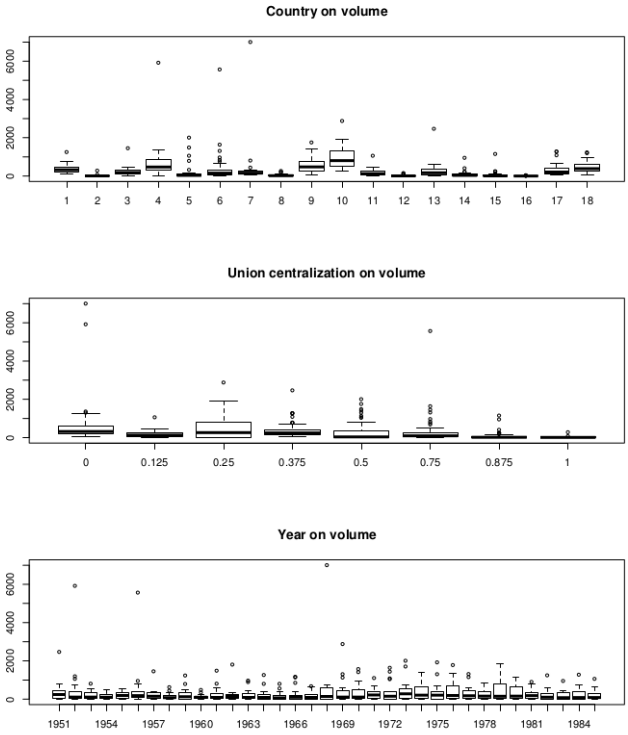

Figure 3: Boxplots for categorical variables

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	16.5621	3.9511	4.19	0.0000
sqrt(unemployment)	-0.7827	0.5592	-1.40	0.1621
inflation	0.3947	0.0725	5.44	0.0000
sqrt(prep)	0.1671	0.5581	0.30	0.7648
as.factor(country)2	-14.6893	1.8878	-7.78	0.0000
as.factor(country)3	-3.3080	1.9612	-1.69	0.0922
as.factor(country)4	7.0721	1.9678	3.59	0.0004
as.factor(country)5	-9.2798	1.8862	-4.92	0.0000
as.factor(country)6	-2.6614	2.0088	-1.32	0.1857
as.factor(country)7	-3.9433	1.8672	-2.11	0.0351
as.factor(country)8	-12.5544	1.8758	-6.69	0.0000
as.factor(country)9	3.7286	2.5731	1.45	0.1478
as.factor(country)10	11.5035	1.9093	6.03	0.0000
as.factor(country)11	-7.2261	1.9213	-3.76	0.0002
as.factor(country)12	-13.8945	1.8875	-7.36	0.0000
as.factor(country)13	-5.9416	1.9600	-3.03	0.0025
as.factor(country)14	-12.1865	1.9284	-6.32	0.0000
as.factor(country)15	-13.5009	1.9079	-7.08	0.0000
as.factor(country)16	-17.1749	2.1118	-8.13	0.0000
as.factor(country)17	-2.6229	1.8673	-1.40	0.1606
as.factor(country)18	2.5133	1.9945	1.26	0.2081

Table 6: Linear model fit

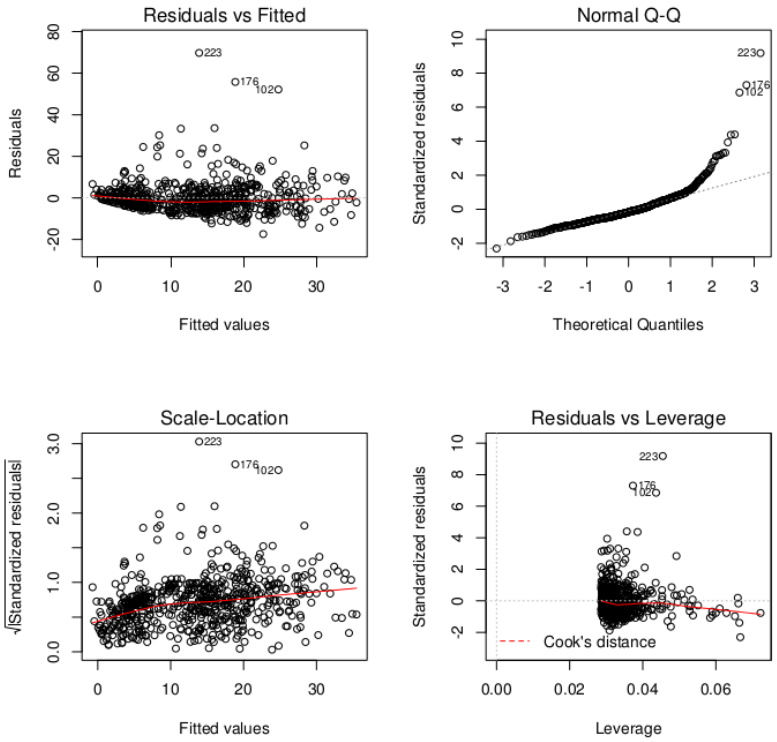

Figure 4: Boxplots for categorical variables

Question 7

Figure 7 shows scatterplots of data sets where observations belong to two different groups.

a) Suppose you don't know the groups (you get a scatterplot where all the dots are black). Could you use linear regression to separate them? Discuss the situation for each panel.

b) Consider the panel on the right. How could you construct a classification rule (using linear regression) and check how good your classification rule is using the same data set?

c) Do you think some other model, rather than linear regression, is better to classify the data? Discuss the situation for each panel.

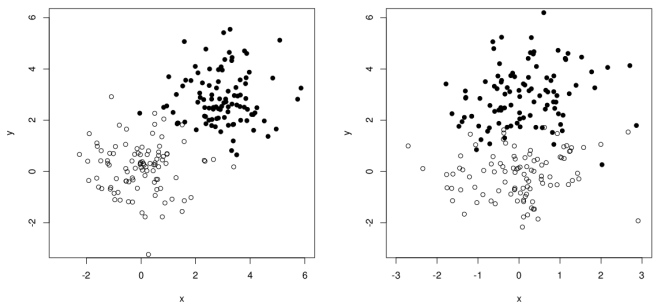

Figure 5: Scatterplots for simulated data