MVE190/MSG500: Linear Statistical Models
Time: 08:30-12:30, Date: 2013-12-19

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 model where the regression line is expected to pass through the origin

$$y_i = \beta x_i + \epsilon_i$$

a) Compute the least squares estimate for β.

b) Derive the variance of $\hat{\beta}$ and $\hat{y}_i$ for this model.

c) Suppose now that the $x_i \geq 0$. Define

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i}, \quad \text{and} \quad \hat{\beta}_2 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}.$$

That is, $\hat{\beta}_1$ and $\hat{\beta}_2$ are two different estimates of β. Show that $\hat{\beta}_1$ and $\hat{\beta}_2$ are unbiased estimates for β.

d) Compare the variances of $\hat{\beta}_1$ and $\hat{\beta}_2$.

e) How do you interpret $\hat{\beta}_1$ and $\hat{\beta}_2$? Which one is a better estimate of β?

Question 2

The following table shows the occurrences of rare words in James Joyce's *Ullyses*.

Number of occurrences	Number of words
1	16432
2	4776
3	2194
4	1285
5	906
6	637
7	483
8	371
9	298
10	222

a) Assume a word in Joyce's vocabulary will appear x number of times according to a Poisson model with mean parameter λ. Estimate λ for the model. Does it predict the counts accurately? Does it seem to be adequate?

b) Consider now that the counts occur according to a Poisson model with parameter λ where

$$\log(\lambda) = \beta_0 + \beta_1 x$$

Figure 1 I shows a scatter plot on the data, the fit from the simple Poisson model from a) and two different GLMs, one for (exactly) the equation above and one for a transformed version of x.

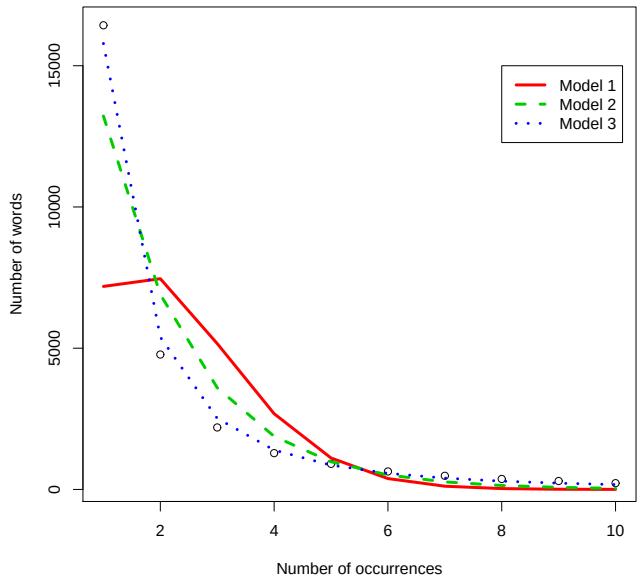

Figure 1: Scatter plot of number of occurrences vs number of counts with fitted model

Figures 2 and 3 show the residual plots for the GLMs. Can you identify which one is which? Comment on the fit of the models with respect to the basic assumptions in generalized linear models.

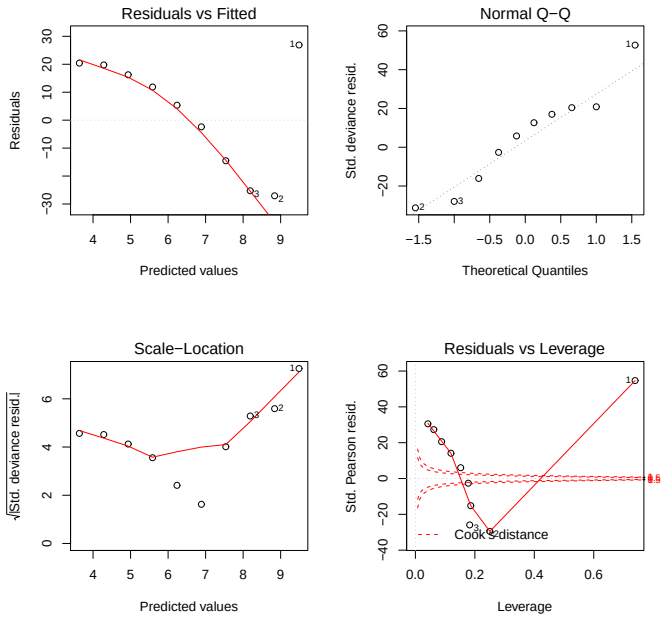

Figure 2: Residual plots from the regression fit.

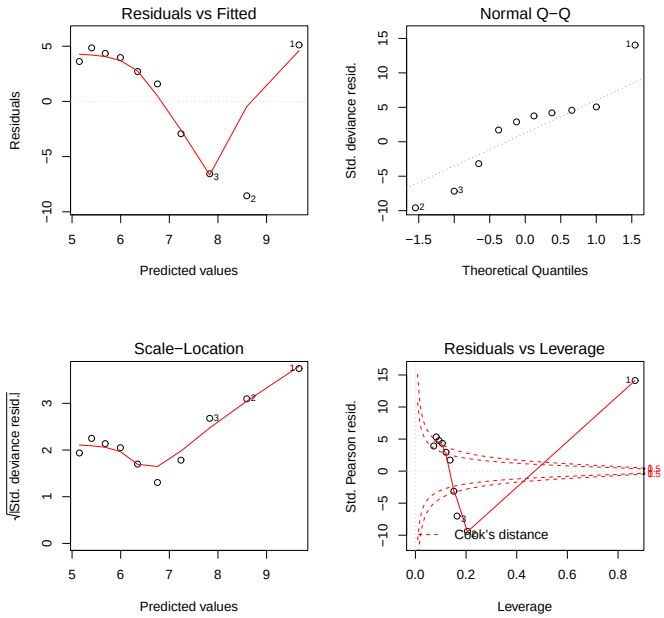

Figure 3: Residual plots from the regression fit.

Question 3

Consider the simple linear regression model

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

Assume that

$$\mathrm{Var}(\epsilon_i) = \begin{cases} \sigma^2 & \text{with probability } 1-\delta \\ \tau^2 & \text{with probability } \delta \end{cases}$$

a) Derive the variance for the OLS estimate of β_1.

b) Based on the result in a), how would you expect the least squares estimate to behave for small perturbations of the data?

c) For a data set $x_1, x_2, \ldots, x_n$, the mean and the median can be defined as

$$\text{mean:} \quad \min_m \left\{ \sum_{i=1}^{n} (x_i - m)^2 \right\}, \quad \text{median:} \quad \min_m \left\{ \sum_{i=1}^{n} |x_i - m| \right\}$$

The usual least squares (OLS) can be thought of as the regression analogue to the mean, while the least absolute deviation (LAD) regression

$$\min_{\beta_0,\beta_1} \left\{ \sum_{i=1}^{n} |y_i - (\beta_0 + \beta_1 x_i)| \right\}$$

can be thought of the regression analogue to the median. The following table shows the measurements of oxygen (O_2) and carbon dioxide (CO_2) in the pouches of 23 potoroos (a marsupial).

Animal	O_2	CO_2
1	20	1
2	19.6	1.2
3	19.6	1.1
4	19.4	1.4
5	18.4	2.3
6	19	1.7
7	19	1.7
8	18.3	2.4
9	18.2	2.1
10	18.6	2.1
11	19.2	1.2
12	18.2	2.3
13	18.7	1.9
14	18.5	2.4
15	18	2.6
16	17.4	2.9
17	16.5	4
18	17.2	3.3
19	17.3	3.4
20	17.8	3.4
21	17.3	2.9
22	18.4	1.9
23	16.9	3.9

To test the performance of OLS and LAD another data set was generated where the O_2 level for animal 15 was changed to 10. Figure 4 shows the OLS and the LAD regressions for the above original data set and the OLS and LAD regressions for the altered data set. Can you motivate which one is which? Why do you expect such behaviour from OLS and from LAD?

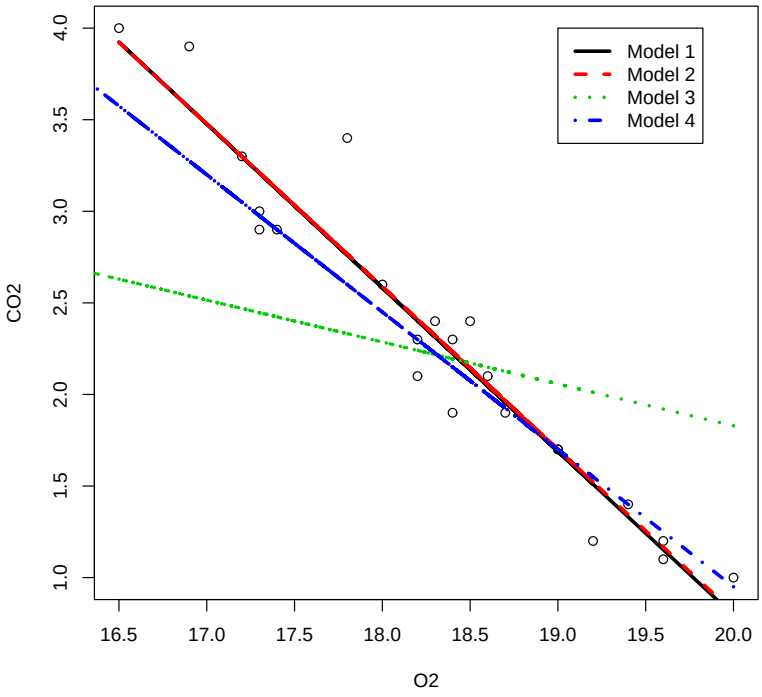

Figure 4: Regression models for the O_2 and CO_2

Question 4

a) Based on survey data, a psychologist runs a regression in which the dependent variable is a measure of depression and the independent variables are marital status. employment status, income, gender and body mass index. Using a regression analysis he finds that people with a higher body mass index are significantly more depressed, controlling for the other variables. Has he shown that being overweight causes depression?

b) In a large organization, the average salary for men is \$47,000 while the average salary for women is \$30,000. A t-test shows that this difference is significant at the 0.001 level. When we control for years of work experience, the p-value for the effect of gender changes to 0.17. What would you conclude?

Question 5

Here we will analyse again the hear disease data set from the lectures. It consists of 312 observations and 12 different variables. Here wi will focus on two of them only, namely, the indicator variable for the presence of heart disease (chd) and cholesterol level (ldl). Cholesterol level is a continuous variable raging form 0.98 to 11.98. I'll transform it to a discrete variable 0/1 by assigning 0 to all observations that have cholesterol level below the mean (4.65), and 1 to all observations that have cholesterol level above the mean; this way cholesterol level becomes an indicator variable for low (those with 0) and high (those with 1) cholesterol level. Below I show a contingency table for this indicator of cholesterol level and presence of heart disease.

	Heart disease		
Cholesterol	Not-present	Present	Total by cholesterol level
Low	131	43	174
High	74	64	138
Total by heart disease	205	107	312

a) Could you suggest a model for the counts in this table using the Multinomial distribution?

b) Assume now that the counts the 4 cells, μ_{ij} for $i, j = 0, 1$, come from a Poisson distribution. A possible model for the logarithm of the counts, the so called log-linear model is

$$\log(\mu_{ij}) = \eta + \alpha_i + \beta_j, \quad i, j = 0, 1.$$

Where α corresponds to cholesterol level and β to heart disease. How do you interpret this model?

c) The p-value for the Pearson statistic where the expected values are computed under this model is less than 10^{-5}. What does this mean in terms of heart disease for people with high cholesterol level?

d) The estimated coefficients for the above model are

Parameter	Value
η	4.298
α_0	0.1159
α_1	-0.1159
β_0	0.3250
β_1	-0.3250

What are the odds of having heart disease versus not having it?

Question 6

The 'Detroit' data set comprises 13 observations and 14 variables for the city of Detroit between years 1961 and 1973 as explained below.

FTP - Full-time police per 100,000 population
UEMP - % unemployed in the population
MAN - number of manufacturing workers in thousands
LIC - Number of handgun licences per 100,000 population
GR - Number of handgun registrations per 100,000 population
CLEAR - % homicides cleared by arrests
WM - Number of white males in the population
NMAN - Number of non-manufacturing workers in thousands
GOV - Number of government workers in thousands
HE - Average hourly earnings
WE - Average weekly earnings
HOM - Number of homicides per 100,000 of population
ACC - Death rate in accidents per 100,000 population
ASR - Number of assaults per 100,000 population

We will use HOM as the response variable and the first 11 variables as predictors (that is, we won't use ACC or ASR). Figure 5 shows some scatter plots for some predictors and the response.

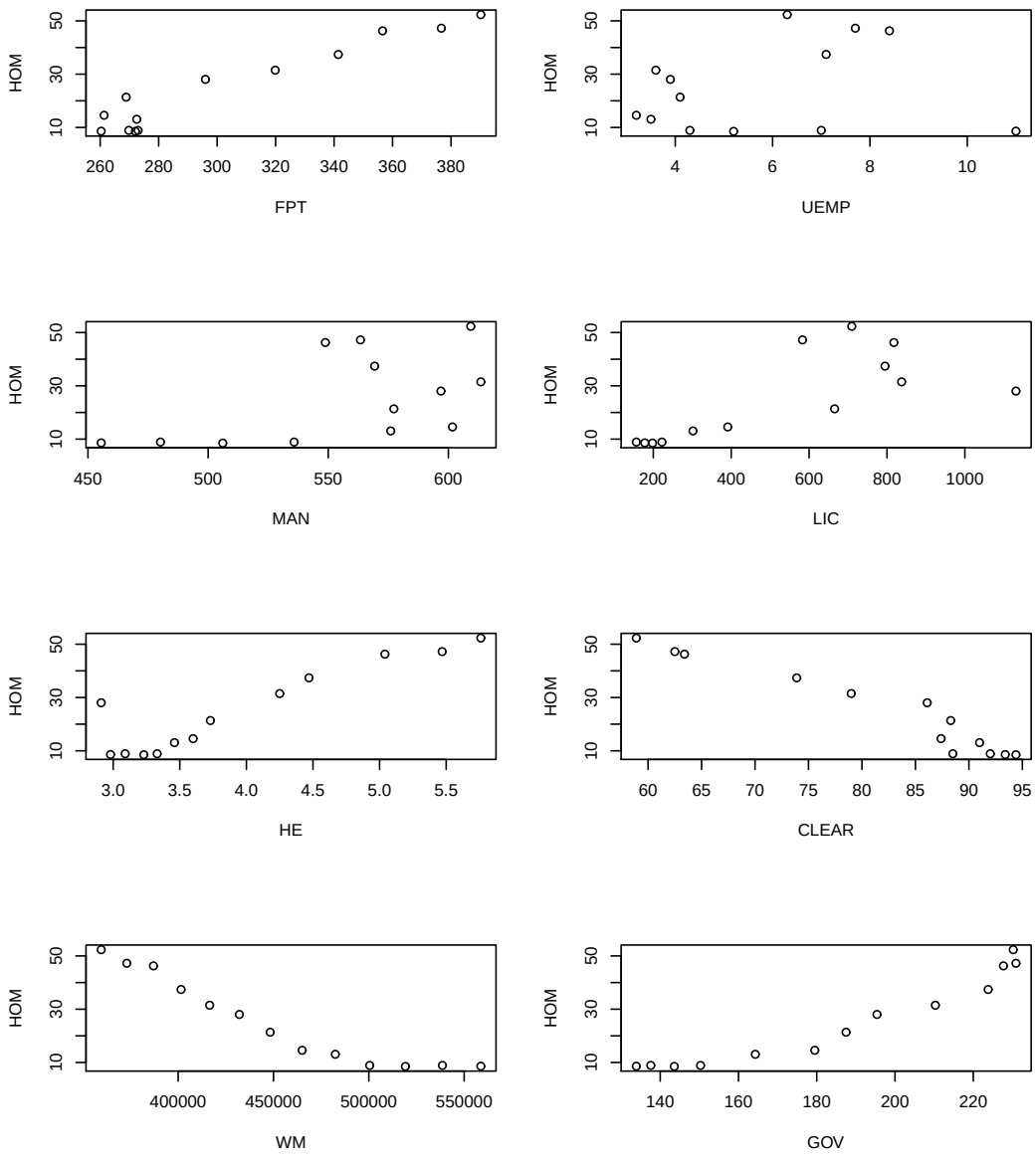

Figure 5: Scatterplots for Detroit data set

a) Does the data seems suitable for a linear regression analysis? Are transformations required?

b) Below are the results for a regression model including the first 11 predictors. The correlation matrix for the variables and the diagnostic plots for the model are also included. Comment on the model overall. What can you say about HOM as a function of the other variables? Interpret the model.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-91.5257	31.8485	-2.87	0.2132
FTP	0.0241	0.0105	2.28	0.2626
UEMP	0.5725	0.1926	2.97	0.2066
MAN	-0.0586	0.0163	-3.60	0.1723
LIC	0.0214	0.0017	12.41	0.0512
GR	-0.0033	0.0013	-2.61	0.2332
CLEAR	-0.0827	0.0616	-1.34	0.4076
WM	0.0001	0.0000	2.24	0.2668
NMAN	0.0496	0.0172	2.88	0.2127
GOV	0.1812	0.0608	2.98	0.2062
HE	-6.0965	1.8919	-3.22	0.1916
WE	0.3052	0.0524	5.83	0.1082

Residual standard error: 0.2998 on 1 degrees of freedom
Multiple R-squared: 1, Adjusted R-squared: 0.9997
F-statistic: 3259 on 11 and 1 DF, p-value: 0.01366

	Int	FTP	UEMP	MAN	LIC	GR	CLEAR	WM	NMAN	GOV	HE	WE
Int	1.00	0.05	0.56	0.67	-0.60	0.55	-0.58	-0.97	-0.56	-0.87	0.74	-0.82
FTP	0.05	1.00	0.11	0.38	-0.37	0.19	0.05	-0.17	-0.17	-0.19	0.28	-0.37
UEMP	0.56	0.11	1.00	0.79	-0.28	0.24	0.04	-0.70	-0.72	-0.43	0.33	-0.33
MAN	0.67	0.38	0.79	1.00	-0.49	0.61	-0.37	-0.74	-0.75	-0.66	0.72	-0.71
LIC	-0.60	-0.37	-0.28	-0.49	1.00	-0.57	0.54	0.60	0.12	0.59	-0.45	0.71
GR	0.55	0.19	0.24	0.61	-0.57	1.00	-0.53	-0.50	-0.33	-0.68	0.76	-0.72
CLEAR	-0.58	0.05	0.04	-0.37	0.54	-0.53	1.00	0.39	0.05	0.60	-0.53	0.67
WM	-0.97	-0.17	-0.70	-0.74	0.60	-0.50	0.39	1.00	0.62	0.85	-0.72	0.79
NMAN	-0.56	-0.17	-0.72	-0.75	0.12	-0.33	0.05	0.62	1.00	0.28	-0.44	0.34
GOV	-0.87	-0.19	-0.43	-0.66	0.59	-0.68	0.60	0.85	0.28	1.00	-0.86	0.91
HE	0.74	0.28	0.33	0.72	-0.45	0.76	-0.53	-0.72	-0.44	-0.86	1.00	-0.93
WE	-0.82	-0.37	-0.33	-0.71	0.71	-0.72	0.67	0.79	0.34	0.91	-0.93	1.00

Table 1: Correlation matrix for predictors

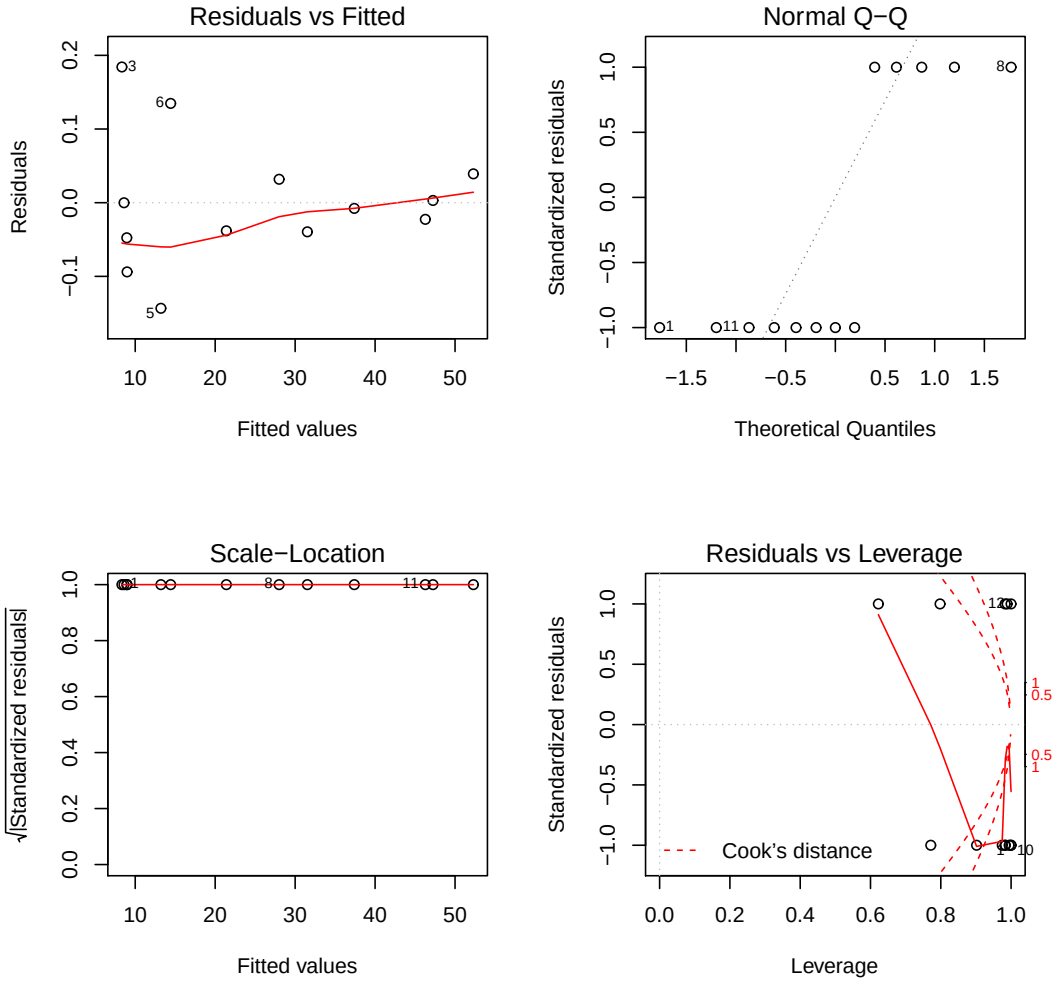

Figure 6: Diagnostic plots

c) Which methods for model selection would be suitable for this data set? Why?

d) After LOOCV a model including FTP, CLEAR and GOV is selected. The results of a regression are included below. Comment on the results.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-19.6364	38.9204	-0.50	0.6260
FTP	0.1142	0.0616	1.85	0.0966
CLEAR	-0.3269	0.2414	-1.35	0.2086
GOV	0.1971	0.0391	5.05	0.0007

Residual standard error: 2.242 on 9 degrees of freedom
Multiple R-squared: 0.986, Adjusted R-squared: 0.9813
F-statistic: 210.7 on 3 and 9 DF, p-value: 1.184e-08

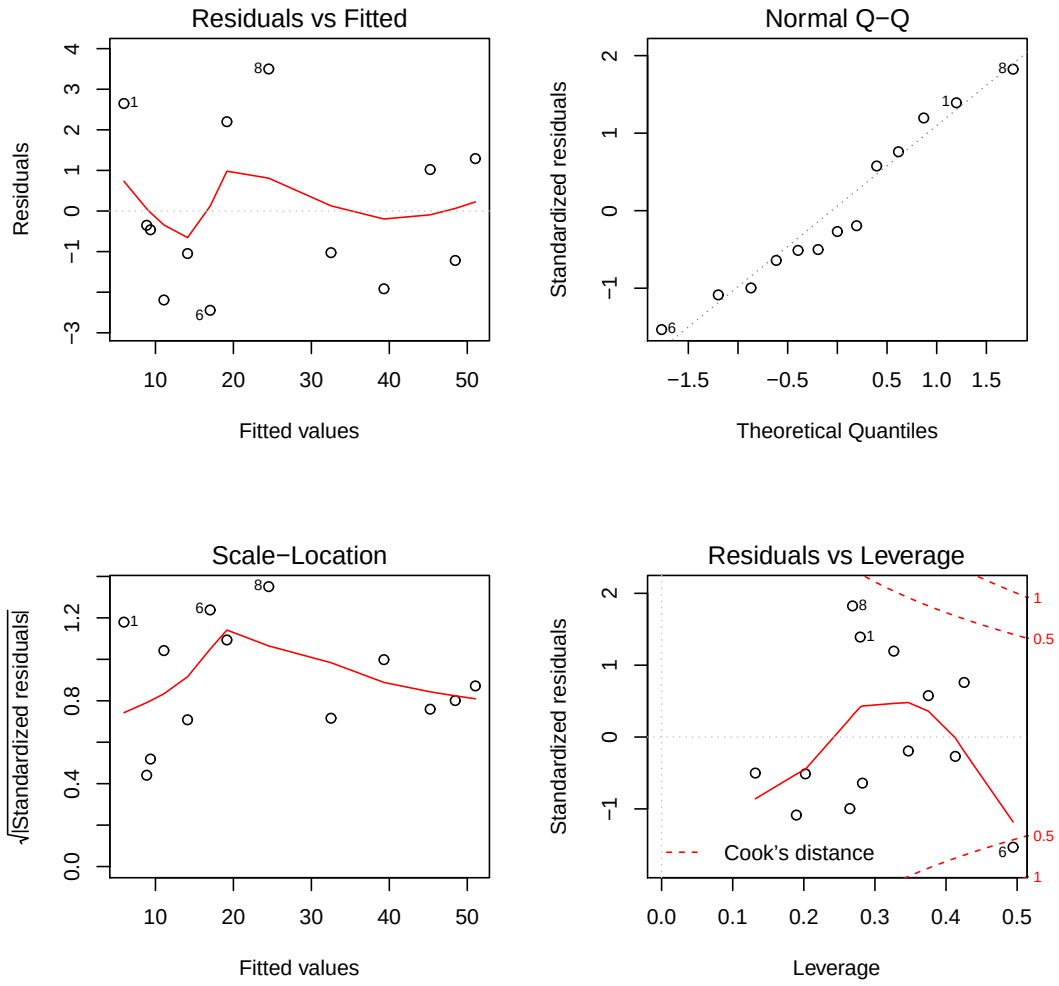

Figure 7: Diagnostic plots

Question 7

The California housing data set contains the median prices as well as location (given by latitude and longitude coordinates) of 20640 houses in California. It is well know that the location is among the most important variables for the price of a house. Below I use CART to construct areas in California where the median prices for houses are similar.

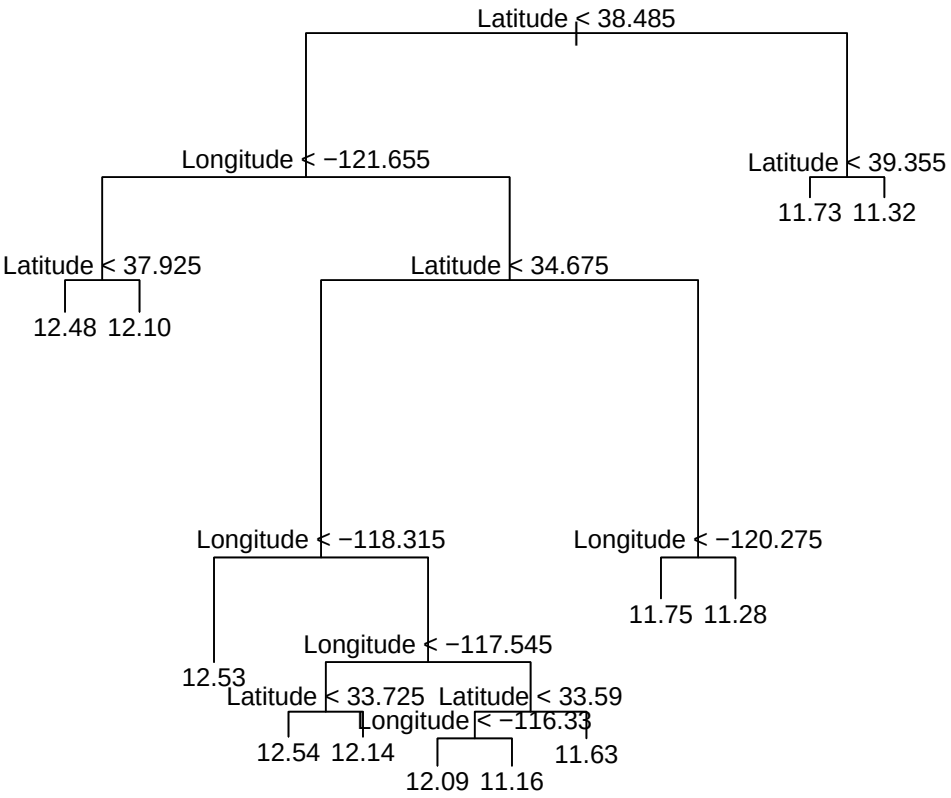

Figure 8: Regression tree

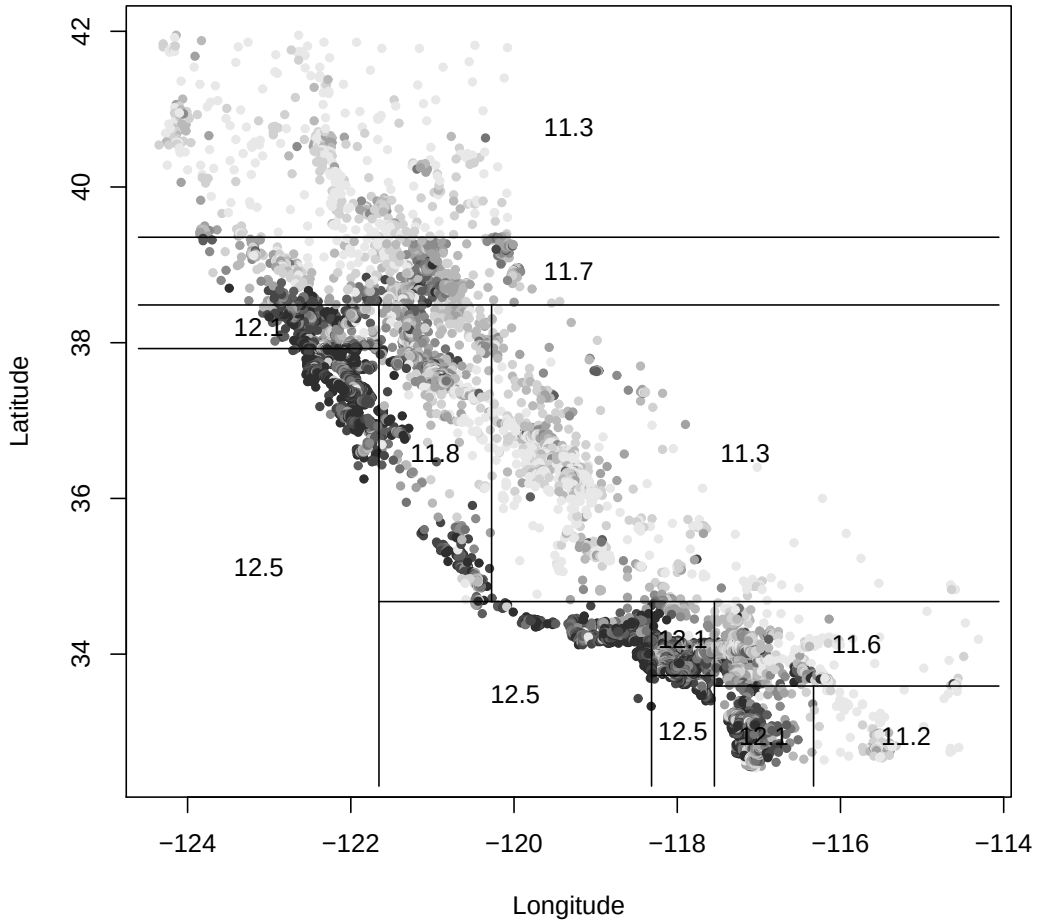

Figure 9: Regression tree

a) Explain how was it done and interpret the results.

b) How does the above analysis compares to regular linear regression?

c) Could you think of another way of constructing areas with similar prices for houses in California?