Exercise week 5

Gherardo Varando, gherardo.varando@math.ku.dk

18/12/2019

Some artificial experiments

In this exercise you are asked to simulate artificial data from some models and then fit the appropriate (and not so appropriate) regressions.

Exercise 1 We study here the simple linear regression using artificially created data.

Ex 1.1 Generate some (n = 50) data from the following model

$$X \sim N(0, \sigma_1^2)$$

$$Y|X = x \sim N(ax + b, \sigma_2^2)$$

And plot the observations in a scatter plot. Which one is the true regression function? Plot the true regression function in red in the same plot together with the observations. (Choose freely the values of a, b, σ_1, σ_2)

Ex 1.2 Fit now a linear regression using the `lm` function. Plot the fitted regression line on top of the previous plot and using a different color (e.g. blue)

Ex 1.3 Use the function `summary` to obtain informations on the coefficients of the model.

Ex 1.4 Repeat from point 1.1 but setting the intercept b in the true model to 2. Check now if in the fitted linear regression we reject $H_0 : Intercept = 0$ (at a level 0.05), play with the σ_2 value and the sample size to see when we reject and not reject the null hypothesis that the intercept is equal to 0.

Ex 1.5 Fit now a regression model without intercept, check the documentation of `formula` to discover how to remove the intercept form the model. Compute AIC, BIC for the models with and without intercept. Perform the F-test. Comment the results.

Exercise 2 We now look at polynomial regression

Ex 2.1 Generate artificial data from the model:

$$X \sim N(0,1)$$

$$Y|X = x \sim N(x^2 - x + 1, 2)$$

In particular obtain a data set of size $n = 50$. Plot also the true regression function.

Ex 2.2 Fit a simple linear regression model $\mathbb{E}(Y|X = x) = \beta_0 + \beta_1 x$ to the data generated in 2.1. Plot the fitted line on top of the scatter plot as usual.

Ex 2.3 Plot the predictor variable vs the residuals and the Q-Q plot of the residuals vs the normal quantiles (`qqnorm` and `qqline` functions), comment the plots.

Ex 2.4 Fit now the true degree 2 polynomial model (remember that it is still a linear model and we can use the function `lm`). Plot the result in the same graph with a different color, you can also add a legend.

Ex 2.5 As in 2.3 obtain also for the polynomial regression model the predictor-residuals plot and the Q-Q plot vs the normal quantiles.

Ex 2.6 Perform model selection between the simple linear regression of point 2.2 and the polynomial regression in point 2.4. Use the log-likelihood ratio test, the F-test (both with `anova`). Moreover perform model selection also using AIC and BIC score (`AIC, BIC`)

Ex 2.7 Try to fit now a polynomial of higher degree (e.g. 3,4,5,...). Perform also here model selection. In particular plot the AIC (or BIC) score as a function of the polynomial degree. Plot also the log-likelihood as a function of the polynomial degree. What can you observe ? What is the difference between the pure log-likelihood and the AIC and BIC scores ?

Exercise 3 We now study non-linear regression.

Ex 3.1 As in exercise 1 and 2 generate $n = 50$ observations from the following model

$$X \sim N(0, \frac{1}{4})$$

$$Y|X = x \sim N(e^{-3x} + 2x, 2)$$

Ex 3.2 Fit a simple linear model $\mathbb{E}(Y|X = x) = \beta_0 + \beta_1 x$, and polynomial regression models up to degree 5.

Ex 3.3 Perform model selection of the previous models using AIC and BIC

Ex 3.4 Check the residuals distribution and the plot the predictor observations vs the residuals. Comment.

Ex 3.4 Now fit the true model $\mathbb{E}(Y|X = x) = \beta_0 + \beta_1 x + \exp(\beta_2 x)$ with Gaussian noise. Try to implement manually the residual sum of squares and minimize it using the `optim` function, similarly implement the log-likelihood function and maximize it with `optim`. Compare the results obtained using `optim` with the coefficients obtained with the appropriate function `nls` that fit non linear regression models with additive Gaussian noise using least square.

Wine quality dataset

We study here the wine quality data (`https://archive.ics.uci.edu/ml/datasets/Wine+Quality`). You can obtain the `winequality-red.csv` file from Absalon.

> The two datasets are related to red and white variants of the Portuguese ”Vinho Verde” wine.

We study here just the red wine data.

Exercise 4 Load the wine quality data with

```
wines <- read.csv("winequality-red.csv", sep =";")
```

Ex 4.1 Fit a linear regression model using all the regressors. Use the function `summary`, based on the results of the t-test which are the important regressors?

Ex 4.2 Use forward stepwise selection with the AIC score to select the relevant covariates. That is, start with the model that includes just the intercept term and iteratively add the variable that obtain a better AIC score (the lower AIC the better). (hint: use the `update` function that updates linear models).

Ex 4.3 A different model selection method is described in AoS, the Zheng-Loh Model Selection Method:

1. Fit the full model and let $W_j = \hat{\beta}_j / \hat{se}(\hat{\beta}_j)$ be the Wald statistic for $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$
2. Order the statistics from largest to smallest in absolute value

$$|W_(1)| \geq |W_(2)| \geq \ldots \geq |W_(k)|$$

3. Let $\hat{j}$ the value of j that minimizes

$$\mathrm{RSS}(j) + j\hat{\sigma}^2 \log(n)$$

Where $\mathrm{RSS}(j)$ is the residual sums of squares from the model using the j terms that corresponds to the largest absolute Wald statistics.

4. The final model is the model using the $\hat{\jmath}$ terms with the largest absolute Wald statistics.

Perform the Zheng-Loh model selection for the wine quality regression.

Exercise 5 In this exercise we consider the binary classification problem obtained from the wine quality dataset. You can download the file winequality-red.csv from Absalon.

Load the data and transform the quality variable to a binary factor with the following code:

```
wines <- read.csv("winequality-red.csv", sep =";")
good <- wines$quality > 5
wines$quality <- "bad"
wines[good, "quality"] <- "good"
wines[,"quality"] <- as.factor(wines[, "quality"])
```

Ex 5.1 Fit a logistic regression model using all the other variables in the dataset as predictors.

Ex 5.2 Implement a forward feature selection based on the AIC or the BIC score. In particular start with the logistic regression model including just the intercept and iteratively add the feature that more decrease the AIC (or BIC) score until no decrease in the score is possible. You can fine useful the function update. Try also a forward feature selection based on log-likelihood (logLik). Be careful that for AIC and BIC lower value of the score is to prefer while for log-likelihood the larger the better.

Ex 5.3 Can you explain why we should not use just the log-likelihood in model selection?

Ex 5.4 Observe the output of the call predict(model) where model is one of the above logistic regression model fitted with the gml function. This is not the good or bad value of the class variable. Any idea on what it is? How can you transform it into the bad/good value ? Try to write a function in this case that transform the output of predict into the class value (you can also try to write a general function for generalized linear models or for logistic regression but it can be more complicated). (hint: use the binomial() call to obtain information on the link and the inverse link function, e.g. binomial()$linkinv gives you the inverse link function)

Ex 5.5 Compute the model accuracy over the data set, that is the proportion of correctly classified observations.

CORIS data

We study here the CORIS data set (from an example of AoS). You can download the data file coris.dat from Absalon.

To load the data set you can use the following

```
coris <- read.table("coris.dat", skip = 4, sep = ",",
                    col.names = c("row.names", "sbp", "tobacco",
                                  "ldl", "adiposity",
                                  "famhist", "typea", "obesity",
                                  "alcohol",
                                  "age", "chd"))[,-1]
```

A quote from AoS:

> The Coronary Risk-Factor Study (CORIS) data involve 462 males between the ages of 15 and 64 from three rural areas in South Africa, (Rousseauwet al. (1983)). The outcome Y is the presence (Y = 1) or absence (Y = 0) of coronary heart disease. There are 9 covariates: systolic blood pressure, cumulative tobacco (kg), ldl (low density lipoprotein cholesterol), adiposity, famhist (family history of heart disease), typea (type-A behavior), obesity, alcohol (current alcohol consumption), and age.

Exercise 6

Ex 6.1 Use backward stepwise selection for logistic regression, use AIC as score and summarize your results.

Ex 6.2 Fit the complete model for logistic regression, that is using all the variables in the data set. What is curious about the coefficients? Which coefficient is a very important indicator of coronary risk? Comment the results.