Statistical models
A.Y. 2013/14
First set of exercises.

Examples of theoretical questions that might be asked in the written test. *[The concepts necessary to answer some of the questions have not been introduced yet in the lectures, but can be found in the book by Faraway 'Practical Regression and Anova using R', available on the Web site]*

1. Describe a procedure to compute a confidence interval for b in the simple regression model $y_i = a + bx_i + e_i$; specify which assumptions are needed about the error terms e_i. *[If possible, arrive at the exact formulae to be used.]*

2. State the Gauss-Markov theorem (including its exact assumptions), describing why it supports the use of least square estimation in linear models.

3. The solution of the least square problem in linear models is based on an orthogonal projection. Explain why the orthogonal projection necessarily satisfy the least square problem; show which is the form of the orthogonal projection for the linear model $Y = X\beta + E$; what does it mean that a linear operator in $\mathbb{R}^n$ is an orthogonal projection?

4. Write down the matrix $(X^t X)$ when we fit a simple regression model to the data points $(x_i, y_i),\ i = 1 \ldots n$. Compute its inverse and arrive at the least-square parameter estimates.
It is advised, to keep the computations simple, to introduce the quantities

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,\ \sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2,$$
$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,\ \sigma_{x,y} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}.$$

5. Explain what is the meaning of the quantity R^2 in linear models. Show that $\sum_{i=1}^{n}(y_i - \bar{y})^2$ can be split as the sum of two sums of squares, and explain their intepretation.

6. After having estimated the parameters $(\hat{\beta}, \hat{\sigma^2})$ of a linear model $Y = X\beta + E$, we perform another experiment (independent of those previously analysed) with a different value of the predictor variables $X_1, \ldots, X_p$; the new value can be written as a row vector $x = (1, x_1, \ldots, x_p)$.

(a) Show that $x\hat{\beta}$ is a reasonable predictor of the value that will be obtained of the response variable Y.

(b) Compute[1] $V(x\hat{\beta})$.

(c) Assume that there is only 1 independent variable (a simple regression) i.e.

$$\beta = \begin{pmatrix} a \\ b \end{pmatrix} \qquad X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

and $x = (1\ t)$. Complete the computations of $V(x\hat{\beta})$, showing that the variance increases, the further away is t from the sample mean $\bar{x}$.

7. Let $Y = X\beta + \varepsilon$, where ε are Gaussian (=normal) random variables with $\mathbb{E}(\varepsilon) = 0$, $\mathrm{Cov}(\varepsilon) = W$, where W is a known positive definite matrix[2] X is a constant (known) matrix, while β is the vector of (unknown) parameters.

   (a) Show that the maximum likelihood estimate of β is the point of minimum of $(Y - X\beta)^t W^{-1}(Y - X\beta)$.

   (b) Write explicitly this minimum problem in the case when W is a diagonal matrix with positive entries σ_i^2.

   (c) Show that this can be phrased as a problem of minimum distance from a subspace in the Euclidean (Hilbert) space $\mathbb{R}^n$ with scalar product $(u, v) = \langle W^{-1/2}u, W^{-1/2}v\rangle$ where $\langle\cdot,\cdot\rangle$ is the standard scalar product, $W^{1/2}$ is the unique [3] positive definite matrix such that $W^{1/2}W^{1/2} = W$ and $W^{-1/2}$ is its inverse.

   (d) Find the minimum point through an orthogonal projection (relatively to the scalar product $(\cdot,\cdot)$), showing that the estimate of β is

$$\hat{\beta} = (X^t W^{-1} X)^{-1} X^t W^{-1} Y.$$

8. We organize an experimental treatment, in which a qualitative variable X_1 can have values 1 and 2, and another qualitative variable X_2 can have values A, B and C. For each combination of the two qualitative variables, 2 observations of the quantitative variable Y are taken. We have thus a total of 12 observations.

[1] It is enough writing the formula without explicit computations.

[2] Look at my notes (in Italian) or at any book of probability theory for the density function of a multivariate normal with covariance matrix $(n \times n)$ W. Otherwise analyse the simpler case where W is a diagonal matrix with positive entries.

[3] its existence and uniqueness can be proved through the diagonalization of W.

We assume that the value of $Y_{i,j,k}$ ($i = 1, 2$ is the value of Z_1, $j = A, B, C$ is the value of Z_2, $k = 1, 2$ represents the observation) is given by the formula

$$Y_{i,j,k} = \mu_i + \nu_j + \varepsilon_{i,j,k}$$

where μ_i and ν_j are some numbers (to be estimated) while $\varepsilon_{i,j,k}$ represent the errors; they are independent and satisfy $\mathbb{E}(\varepsilon_{i,j,k})$ and $V(\varepsilon_{i,j,k}) = \sigma^2$.

Write down the corresponding model matrices X and the matrix $(X^t X)^{-1}$ in the additive case. Write down expected value and variance for the estimators in the resulting model[4]

9. Assume that we obtain a dataset that we model with a linear model $Y = X\beta + E$ with $E \sim N(0, \sigma^2 I)$ and X an $(n \times 4)$ matrix.

   (a) Explain how a test of the hypothesis $\beta_1 = 0$ is performed[5].

   (b) Explain why a test of the hypothesis $H_0 : \beta_1 = 0$ against the alternative $H_1 : \beta_1 \neq 0$ has a different rejection region than the test (at the same level α) of H_0 against the alternative $H_1 : \beta_1 > 0$.

   (c) Explain why a test of the hypothesis $H_0 : \beta_1 = 0$ may give different result if we use the linear model described above, or if we use a reduced matrix X to which we deleted the last column.

10. Running `lm` in R on the dataset `cathedral` produced the following output

```
Call:
lm(formula = cathedral$x ~ cathedral$y + cathedral$style)

Residuals:
    Min      1Q  Median      3Q     Max
-15.184  -8.107  -2.115   8.014  23.360

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      34.96916    9.41432   3.714  0.00121 **
cathedral$y       0.10058    0.02259   4.452  0.00020 ***
cathedral$styler -8.34535    5.03843  -1.656  0.11185
---
```

[4] it is required just to write the ingredients for the computation, not an explicit formula that would be very cumbersome.

[5] It is not necessary to explain the computations, but just to explain the method, referring also to the output in R

```
Residual standard error: 11.33 on 22 degrees of freedom
Multiple R-squared:  0.474,Adjusted R-squared:  0.4262
F-statistic: 9.914 on 2 and 22 DF,  p-value: 0.0008521
```

(a) Compute a 95% and 99% confidence interval for β_1 and β_2[6]

(b) From the output in R, was it possible to immediately tell whether these confidence intervals would include 0?

(c) Knowing that y is a quantitative variable, while `style` is a qualitative variable with the 2 values g (used as reference) and r, describe the estimated formula to predict x.

11. Explain what is a maximum likelihood estimate, and a maximum likelihood ratio test, giving an example of each.

12. Assume we observe a quantitative variable Y and a qualitative one X that has 3 possible values, say A, B and C. We assume that $Y = \mu_s + N(0, \sigma^2)$ where s stands for A, B or C.

    We wish to test, allowing for a probability α of an error of the first kind, whether there is any difference between μ_A, μ_B and μ_C.

    Explain why it is not appropriate to do so by first comparing μ_A and μ_B ; then μ_B and μ_C; finally μ_A and μ_C all with a t-test at the level α.

    Explain[7] how this is tested in an analysis of variance, and show which is the matrix X when this is written in the linear model syntax $Y = X\beta + E$.

13. Which are the exact assumptions in the linear regression $y_i = a + bx_i + \varepsilon_i$? How can these be checked by analysing the residuals of the regression? Which ones can be corrected by using the logarithmic transformation, i.e. using $z_i = \log(y_i)$ in place of the original values y_i, assuming of course that they must be positive?

14. Consider a linear model with response variable Y and two predictor variables, X_1 quantitative and X_2 qualitative (with three values, say A, B and C).

[6] The computations require the use of a table (possibly digital) of the t-distribution. You can also approximate this by using a normal distribution, but specify also the procedure that would have been used having a table.

[7] you can just give the idea, referring to the output of lm in R, without giving the details of the formulae

(a) Write down (in a mathematical way) the assumptions of the additive model ($Y \sim X_1 + X_2$ in R).

(b) Give a graphical representation of this model

(c) Which are the (null and alternative) hypotheses that are routinely tested in this model?

(d) Write down (in a mathematical way) the assumptions of the model with interaction ($Y \sim X_1 * X_2$ in R).

(e) Give a graphical representation of this model

(f) Does this model differ (and, if so, how) from performing separate linear regressions of Y on X_1 ($Y \sim X_1$ in R) for each of the subsets $\{X_2 = A\}$, $\{X_2 = B\}$, $\{X_2 = C\}$?

(g) How can we decide whether to choose the model with or without interaction?

15. Which of the following models can be directly written as linear models? Which, after some transformation? *In all cases ε_i represent independent and equidistributed error terms, $\mathbb{E}(\varepsilon_i) = 0$; a, b, $c \ldots$ parameters to be estimated.*

(a) $y_i = \begin{cases} a + b(x_i - 30) + \varepsilon_i & \text{if } x_i < 30 \\ a + c(x_i - 30) + \varepsilon_i & \text{if } x_i > 30 \end{cases}$

(b) $y_i = |a + bx_i + cx_i^2 + \varepsilon_i|$

(c) $y_i = \exp\{a + bx_i + cx_i^2 + \varepsilon_i\}$

(d) $y_i = a\cos(bx_i + c) + \varepsilon_i$

(e) $y_i = ax_i^b(1 + \varepsilon_i)$

(f) $y_i = a + \frac{bx_i}{x_i+1}$

(g) $y_i = ax_iz_i + bx_i^2z_i$.

16. Show that the log-likelihood of a linear model (i.e., the logarithm of the maximum likelihood of the parameters, given the data) multiplied by -2 is (under the assumption of Gaussian errors) equal to $n\log(RSS)$, plus some terms that do not depend on parameters, where RSS is the residual sum of squares.

17. Prove that the following relation holds

$$V(\hat{\varepsilon}_i) = (1 - H_{ii})\sigma^2$$

for the residuals $\hat{\varepsilon}_i = y_i - \hat{y}_i$ where y_i are the observations assumed to satisfy $y = X\beta + \varepsilon$, $\hat{y} = X\hat{\beta}$ are the predicted values and $Cov(\varepsilon) = \sigma^2 I$. H_{ii} is called the *leverage* of the observation i. Explain why the previous relation implies that observation with high leverage will generally correspond to small $|y_i - \hat{y}_i|$.