Avd. Matematisk statistik

KTH Matematik

EXAM IN SF2930 REGRESSION ANALYSIS TUESDAY 14th MARCH 2017, 08.00-13.00.

Examiner: Tatjana Pavlenko, tel. 08-790-8466, email: pavlenko@math.kth.se

Tillåtna hjälpmedel/ Means of assistance permitted : L. Råde & B. Westergren: Mathematics Handbook for Science and Engineering and pocket calculator.

General instructions: You should define and explain all the notations used. Your line of reasoning, assumptions and computations must be clearly written down so that they are easy to follow. You may apply results stated in a part of an exam question to another part of the exam question even if you have not solved the first part. Solutions written in Swedish are, of course, welcome.

The number of exam questions (Uppgift) is six (6). Each question gives maximum six (6) points. 18 points will guarantee a passing result.

The exam results will be announced at the latest on Wednesday the 5^{th} of April, 2017.

Your graded exam paper can be retained at the Student affairs office of the Department of Mathematics during a period of seven weeks after the date of the exam.

Good luck!

Uppgift 1

Consider the linear regression model with k regressors

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{y}$ is $n \times 1$, $\mathbf{X}$ is $n \times p$, $\boldsymbol{\beta}$ is $p \times 1$, $\boldsymbol{\varepsilon}$ is $n \times 1$, and $p = k + 1$.

a) Find the ordinary least squares (OLS) estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$. Make the proper normality assumptions and derive the distribution of $\widehat{\boldsymbol{\beta}}$ under these assumptions.

b) State the test of significance of a single slope parameter β_j and derive the test statistics (t-tests) in the multiple regression setting, $j = 1, \ldots, k$

c) Describe the situations in regression analysis where the assumption of normal distribution is crucial and where it is not (coefficients and mean response estimates, tests, confidence intervals, prediction intervals). Clear motivation must be presented.

Uppgift 2

Consider the linear regression model with k regressors

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{y}$ is $n\times 1$, $\mathbf{X}$ is $n\times p$, $\boldsymbol{\beta}$ is $p\times 1$, $\boldsymbol{\varepsilon}$ is $n\times 1$, and $p = k+1$. Cook's distance measure, denoted by D_i and used for detecting potentially influential observations in regression analysis, is defined as

$$D_i = D_i(\mathbf{X}'\mathbf{X}, pMS_{Res}) = \frac{(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})}{pMS_{Res}}, \quad i = 1, \ldots, n,$$

where $\widehat{\boldsymbol{\beta}}$ is OLS estimator of $\boldsymbol{\beta}$ obtained by using all n observations, $\widehat{\boldsymbol{\beta}}_{(i)}$ is the estimator obtained with point i deleted, $MS_{Res} = SS_{Res}/(n-p)$ and SS_{Res} denotes the residual sum of squares.

a) Show formally that the Cook's D_i depends on both the residual, e_i and the leverage, h_{ii}, and can be expressed as

$$D_i = \frac{r_i^2}{p}\frac{h_{ii}}{1-h_{ii}}, \quad \text{where} \quad r_i = \frac{e_i}{\sqrt{MS_{Res}(1-h_{ii})}}$$

is the studentized residual and h_{ii} is the ith diagonal element of the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

b) Explain how to detect potentially influential data points using Cook's distance measure. Explain further why the representation of D_i given in a) is desirable for assessing the influence. Sketch graphs that support your explanations.

Hint: Use the representation

$$\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)} = \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i e_i}{1-h_{ii}}$$

and recall that $h_{ii} = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$.

Uppgift 3

Suppose that the error component, $\boldsymbol{\varepsilon}$, in the multiple regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$, has mean $\mathbf{0}$ and covariance matrix $\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{\Omega}$, where $\boldsymbol{\Omega}$ is a known $n\times n$ positive definite symmetric matrix and $\sigma^2 > 0$ is a constant (possibly unknown but you do not need to estimate it). Let

$$\widehat{\boldsymbol{\beta}}_{\text{GLS}} = \left(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y}$$

be the generalized least-squares estimator of $\boldsymbol{\beta}$.

a) Show formally that $\widehat{\boldsymbol{\beta}}_{\text{GLS}}$ is obtained as the solution of the problem

$$\text{Minimize}_{\boldsymbol{\beta}}\left[(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Omega}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right].$$

b) Show formally that $\widehat{\boldsymbol{\beta}}_{\text{GLS}}$ is an unbiased estimator of $\boldsymbol{\beta}$ and determine its covariance matrix.

Hint: Use the following general matrix derivatives rules. Let $\boldsymbol{A}$ be $k \times k$ matrix of constants, $\boldsymbol{a}$ be a $k \times 1$ vector of constants and $\mathbf{v}$ be a $k \times 1$ vector of variables. Then the following holds.

$$\text{If} \quad \boldsymbol{z} = \boldsymbol{a}'\mathbf{v}, \quad \text{then} \quad \frac{\partial \boldsymbol{z}}{\partial \mathbf{v}} = \frac{\partial \boldsymbol{a}'\mathbf{v}}{\partial \mathbf{v}} = \boldsymbol{a}.$$

$$\text{If} \quad \boldsymbol{z} = \mathbf{v}'\mathbf{v}, \quad \text{then} \quad \frac{\partial \boldsymbol{z}}{\partial \mathbf{v}} = \frac{\partial \mathbf{v}'\mathbf{v}}{\partial \mathbf{v}} = 2\mathbf{v}.$$

$$\text{If} \quad \boldsymbol{z} = \boldsymbol{a}'\boldsymbol{A}\mathbf{v}, \quad \text{then} \quad \frac{\partial \boldsymbol{z}}{\partial \mathbf{v}} = \frac{\partial \boldsymbol{a}'\boldsymbol{A}\mathbf{v}}{\partial \mathbf{v}} = \boldsymbol{A}'\boldsymbol{a}.$$

$$\text{If} \quad \boldsymbol{z} = \mathbf{v}'\boldsymbol{A}\mathbf{v} \,\text{and}\, \boldsymbol{A} \,\text{is symmetric, then} \quad \frac{\partial \mathbf{v}'\boldsymbol{A}\mathbf{v}}{\partial \mathbf{v}} = 2\boldsymbol{A}\mathbf{v}.$$

Uppgift 4

Explain in detail the idea of ridge- and Lasso regression. Which of this two approaches behaves only as a shrinkage method and which one can directly perform variable selection? Your answer must be formulated in mathematical terms. Provide geometric interpretation of the constraints used in ridge- and Lasso estimation approaches to confirm your answer. Sketch the graph with traces of ridge- and Lasso coefficient estimators as tuning parameter is varied, and explain the difference in trace shapes.

Uppgift 5

a) Suppose that you fit the model $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$ when the true model is actually given by $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$. For both models, assume that $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $Var(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$, $\sigma > 0$. Find the ordinary least squares (OLS) estimator $\widehat{\boldsymbol{\beta}}_1$ of $\boldsymbol{\beta}_1$. Under what conditions is this estimator unbiased? Your answer must be stated in mathematical terms.
Hint: Derive the expected value of $\widehat{\boldsymbol{\beta}}_1$.

b) Regression analysis often utilities the variable selection procedure know as the *all possible regressions* (also called for the *best subsets regression*). Describe thoroughly the steps of the all possible regressions procedure. Specify at least two objective criteria that can be used for the model evaluation, explain how to apply these criteria and motivate why they are suitable for this type of variable selection. Explain advantages and disadvantages of all possible regressions approach to the model building.

Uppgift 6

Assume that $Y \in \text{Be}(\sigma(\boldsymbol{\beta}'\mathbf{x}))$, where $\sigma(\boldsymbol{\beta}'\mathbf{x})$ is the logistic function, $\boldsymbol{\beta}'\mathbf{x} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$ and $\mathbf{x} = (1, x_1, x_2, \ldots, x_p)$, i.e., Y follows a logistic regression.
Let $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)$ be a data set of independent samples, where $y_i \in \{0, 1\}$ and $\mathbf{x}_i = (1, x_{i1}, x_{i2}, \ldots, x_{ip})$. Show that for all real $\boldsymbol{\beta}$, the log likelihood function $l(\boldsymbol{\beta})$ can be written as

$$l(\boldsymbol{\beta}) = -\sum_{i=1}^{n} \ln\left(1 + e^{(1-2y_i)\boldsymbol{\beta}'\mathbf{x}_i}\right).$$