Avd. Matematisk statistik

KTH Matematik

EXAM IN SF2930 REGRESSION ANALYSIS TUESDAY, JUNE 2, 2020, 08.00-13.00.

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

Tillåtna hjälpmedel/ Means of assistance permitted : NONE.

General instructions: You should define and explain all the notations used. Your line of reasoning, assumptions and computations must be clearly motivated and 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.

GOOD LUCK!

Uppgift 1

a) Describe the principle of least-squares and use it to derive the normal equations

$$n\hat{\beta}_0 + (\sum_{i=1}^{n} x_i)\hat{\beta}_1 = \sum_{i=1}^{n} y_i$$
$$(\sum_{i=1}^{n} x_i)\hat{\beta}_0 + (\sum_{i=1}^{n} x_i^2)\hat{\beta}_1 = \sum_{i=1}^{n} x_i y_i.$$

for the simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2),\ i = 1, \ldots, n.$$

Solve the normal equations to obtain the least-squares estimates of β_0 and β_1.

b) Suppose that we have fit the simple (straight-line) regression model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$ but the response variable is affected by the second variable x_2 so that the true regression function is

$$\mathrm{E}(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

Is the least-squares estimator of the slope in the original simple linear regression model unbiased? Your answer must be motivated mathematically.

Uppgift 2

a) Explain the problem of hidden extrapolation in predicting new responses and estimating the mean response at given point $\mathbf{x}_0' = [1, x_{01}, x_{02}, \ldots, x_{0k}]$ in the multiple linear regression. Motivate your explanations by sketching the graph and explain how to detect this problem by using the properties of the hat matrix, $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Recall that the location of the point $\mathbf{x}_0'$ relative to the regressor variable hull is reflected by $h_{00} = \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0$.

b) For the simple linear regression model, show formally that the elements of the hat matrix, $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ are

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{S_{xx}} \quad \text{and} \quad h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}},$$

$i, j = 1, \ldots, n$. Discuss the behavior of these quantities as x_i moves away from $\bar{x}$.

Hint: Recall that

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{S_{xx}} \begin{pmatrix} \sum_{i=1}^n x_i^2/n & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}.$$

Uppgift 3

a) Consider the multiple linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ (in matrix notations). Suppose that $\mathbf{X}'\mathbf{X}$ is in the correlation form, $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues of $\mathbf{X}'\mathbf{X}$, and $\mathbf{T}$ is the corresponding matrix of eigenvectors. Show formally that VIFs, variance inflation factors, are the main diagonal elements of the matrix $\mathbf{T}\boldsymbol{\Lambda}^{-1}\mathbf{T}'$.

b) Explain in mathematical terms the idea of principal-component regression (PCR) and how this approach combats the problem of multicollinearity in the linear regression models.

Uppgift 4

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 $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$, (in matrix notations). For both models, assume that $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$, $\sigma > 0$. 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) Consider a correctly specified regression model with p terms, including the intercept. Make the usual assumptions about the error term $\boldsymbol{\varepsilon}$ and prove that

$$\sum_{i=1}^n \mathrm{Var}(\hat{y}_i) = p\sigma^2.$$

Hint: Recall that $rank(\mathbf{X}) = p$ and that the diagonal elements h_{ii} of the hat matrix $\mathbf{H}$ can be expressed as $\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$, where $\mathbf{x}_i$ is the ith row of $\mathbf{X}$, $i = 1, \ldots, n$.

Uppgift 5

a) Two main sampling procedures for bootstrapping regression estimates are usually referred to as *bootstrapping residuals* and *bootstrapping cases*. Give in mathematical terms the steps of both procedures and specify the difference between these two approaches.

b) Explain how to find a bootstrap estimate of the standard deviation of the estimate of the mean response at a particular point $\mathbf{x}_0$. Explain in mathematical terms how to obtain approximate confidence intervals for regression coefficients through bootstrapping.

Uppgift 6

Ridge regression is a regularization method for the linear model which looks for the vector $\boldsymbol{\beta}$ that minimizes the penalized residual sum of squares,

$$\boldsymbol{\beta}_{\text{Ridge}} = \arg\min_{\boldsymbol{\beta}} \left\{ (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_2^2 \right\},$$

where $\|\boldsymbol{\beta}\|_2^2 = \sum_{j=1}^{p} \beta_j^2$ denotes the squared L_2-norm of $\boldsymbol{\beta}$ and $\lambda \geq 0$ is the regularization parameter. Assume that the $n \times p$ design matrix $\mathbf{X}$ is fixed and the components of $\boldsymbol{\beta}$ are independently distributed as normal random variables with mean 0 and known variance $0 < \tau^2 < \infty$, i.e. the prior knowledge about the vector of coefficients $\boldsymbol{\beta}$ is summarized in terms of the normal *prior*, $\boldsymbol{\beta} \sim N_p(\mathbf{0}, \tau^2 \boldsymbol{I})$. Assume further normal sampling model for the response variable, so that $\mathbf{Y}|\mathbf{X}, \boldsymbol{\beta} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I})$ where $0 < \sigma < \infty$ is a known constant. Show that the ridge regression estimator is the mean vector (and mode) of the *posterior* distribution of $\boldsymbol{\beta}$. Find the relationship between the regularization parameter λ and the variances σ^2 and τ^2.

Hint: The density of k-dimensional normal distribution $N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ ($\boldsymbol{\Sigma}$ assumed to be a positive definite $k \times k$ matrix) is given by

$$f(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^k \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2} (\boldsymbol{y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}) \right),$$

Recall further that the posterior density of $\boldsymbol{\beta}$ is proportional to the likelihood times the prior.