SF2930 Regression analysis

Questions to be considered for the written exam

This document contains a set of assignments and conceptual questions on the topics treated in SF2930 Regression Analysis during the period 3 of 2017. Questions are constructed by Henrik Bosaeus, Alexandre Chotard, Timo Koski, Tatjana Pavlenko and Felix Rios. Six of these questions (or their slightly modified versions) will be selected to constitute the written exam on Tuesday, the 14th of March, 2017, 08.00-13.00. Observe that *Hint* is given after some of the questions; this hint summarizes the formulas which will be provided for this type of question during the exam.

The answers and solutions can be obtained by study of the relevant chapters in the main course textbook, *Introduction to Linear Regression Analysis* by D. Montgomery, E. Peck, G. Vining, Wiley, 5th Edition (2012) (abbreviated in what follows by MPV), other books suggested as a course literature, see

`https://www.math.kth.se/matstat/gru/sf2930/regplanr2017.html`

by similar study of the your own lecture notes and material on the webpage

`https://www.math.kth.se/matstat/gru/sf2930/courselog17.html`

Observe that the derivations presented on the board during the lectures are also topics of the examination. In addition, some proficiency in manipulating basic calculus, probability, linear algebra and matrix calculus is required.

This same set of questions (may be some will be removed and new added) will be valid in the re-exam. Hence we shall NOT provide a solutions manual.

Simple linear regression

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 linear regression model

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

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

2. Derive the estimate of β_1 in the no-intercept model $y_i = \beta_1 x_i + \epsilon_i, i = 1, \ldots, n$, from the least squares criterion, that is to minimize $S(\beta_1) = \sum(y_i - \beta_1 x_i)^2$. Give examples of when such model can be appropriate/inappropriate.

3. Verify the properties of residuals presented in 1.– 5. (see p. 20 MPV).

4. Explain the difference between the confidence interval for estimating the mean response for a given value of the predictor x and the prediction interval for predicting a new response for a given value of the predictor x in the simple linear regression setting. To support your explanations, sketch the graph and describe the relationship between the two confidence bands.

5. In the analysis-of-variance, ANOVA approach to testing the significance of regression, the total variation in a response y is broken down/decomposed into two parts - a component that is due to the regression or model, and a component that is due to random error. Derive this decomposition, use it to explain the construction of the ANOVA table and derive the ANOVA F-test for testing significance of regression.

6. Exercises from MPV: 2.25, 2.27, 2.29, 2.33.

Multiple linear regression

1. (a) State the multiple linear regression model in matrix notations, form normal equations and derive the solution using ordinary least-squares (OLS) estimation approach. State exactly model assumptions under which OLS estimator of the vector of regression coefficients is obtained.

   (b) Show formally that the OLS estimator of the vector of regression coefficients is an unbiased estimator under the model assumption specified in part a). Find the covariance matrix of the vector of estimated coefficients.

2. (a) For the model, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$, (in matrix notations) obtain the 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.

   (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.

3. For the linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ (in matrix notations) where $\boldsymbol{\varepsilon}$ has zero mean, define the error sum of squares as

$$SS_e(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

   For the OLS estimator $\widehat{\boldsymbol{\beta}}$, show that

$$SS_e(\boldsymbol{\beta}) = SS_{Res} + (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}),$$

   where $SS_{Res} = SS_e(\widehat{\boldsymbol{\beta}})$.

4. 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 $h_{00} = \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0$.

5. Exercises from MPV: 3.27, 3.28, 3.29 (*Hint*: Recall that for the hat matrix, $\mathbf{H}$, each element h_{ij} can be expressed as $h_{ij} = [1 \quad x_i](\mathbf{X}'\mathbf{X})^{-1}[1 \quad x_j]'$), 3.32, 3.37, 3.38 (Hint: Recall that $rank(X) = p$ 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 $\mathbf{X}$, $i = 1, \ldots, n$.

Transforms and weighting. Detection of outliers, high leverage observations and influential data points.

1. Define some different types of residuals (for example standardized, studentized or PRESS), specify their properties, and explain how they can be used for detecting outliers.

2. Derive the concept of an influential data point (sketch the graph) and explain how such points can be detected using DFFITS and Cook's distance measure.

3. Cook's distance measure, denoted by D_i and used for detecting potentially influential observations, 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 and $MS_{Res} = SS_{Res}/(n-p)$.

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 $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Explain why this representation of D_i in terms of both the location of the point in x space and the response variables is desirable (for detecting influential points).

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

4. Exercises from MPV: 5.8, 5.14, 5.15 (Hint: For the case of simple linear regression model without intercept, the weighted LS function is given by $S(\beta) = \sum_{i=1}^{n} w_i(y_i - \beta x_i)^2$).

5. Suppose that the error component, $\boldsymbol{\varepsilon}$, in the multiple regression model(Obs! Model is in vector form) $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, 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 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.

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

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

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

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

Multicollinearity

1. Explain in detail (with formulas) the concept of multicollinearity in multiple linear regression models. Describe in detail (with formulas) at least two effects of multicollinearity on the precision accuracy of the regression analyses. Explain why the ordinary LS parameter estimation in multiple regression model is not applicable under strong multicollinearity.

2. Derive in detail at least two diagnostic measures for detecting multicollinearity in multiple linear regression and explain in which way these measures reflect the degree of multicollinearity.

3. Suppose that there are two regressor variables, x_1 and x_2, in the linear regression model. Assuming further that both regressors and the response variable y are scaled to unit length, the model is $y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$, where $\mathrm{E}(\varepsilon_i) = 0$, $\mathrm{V}(\varepsilon_i) = \sigma^2$ and $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$, $i, j = 1, \ldots, n$.

   State the least-squares normal equations in matrix notations and obtain the estimators of β_1 and β_2. Show formally why the strong multicollinearity between x_1 and x_2 results in large variances and covariances for the least-squares estimators of the regression coefficients.

   Hint: Recall that in the unit length scaling, the matrix $\mathbf{X}'\mathbf{X}$ is in the form of correlation matrix and similarly, $\mathbf{X}'\mathbf{y}$ is in the correlation form, that is

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1k} \\ r_{12} & 1 & r_{23} & \cdots & r_{2k} \\ r_{13} & r_{23} & 1 & \cdots & r_{3k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{1k} & r_{2k} & r_{3k} & \cdots & 1 \end{bmatrix} \quad \mathbf{X}'\mathbf{y} = \begin{bmatrix} r_{1y} \\ r_{2y} \\ r_{3y} \\ \vdots \\ r_{ky} \end{bmatrix},$$

   where r_{jl} is the simple correlation between regressors x_j and x_l, and r_{jy} is the simple correlation between the regressor x_j and the response y, $j, l = 1, 2, \ldots, k$. Recall further that in general, for the LS estimator of p-vector $\boldsymbol{\beta}$, $\mathrm{Var}(\hat{\beta}_j) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{jj}$ and $\mathrm{Cov}(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{ij}$, where $(\mathbf{X}'\mathbf{X})^{-1}_{jj}$ and $(\mathbf{X}'\mathbf{X})^{-1}_{ij}$ are diagonal and off-diagonal elements of the the matrix $(\mathbf{X}'\mathbf{X})^{-1}$, respectively, $i, j = 1, \ldots, p$.

4. 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 $\mathbf{T}\boldsymbol{\Lambda}^{-1}\mathbf{T}'$.

Biased regression methods

1. Explain the idea of the ridge regression (in relation to multicollinearity) and define the ridge estimator of the vector of regression coefficients for the linear

model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Show that the ridge estimator is a linear transform of the ordinary LS estimator of regression coefficients and explain why the ridge estimator is also called for shrinkage estimator.

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

3. Explain the idea of the ridge regression and Lasso regression and the difference between these two approaches. Specifically, which of this two approaches behaves only as a shrinkage method and which one can directly perform variable selection? Motivate your explanations by sketching the graph with traces of ridge- and Lasso coefficient estimators as tuning parameter is varied, and explain the difference in trace shapes.

4. Consider the multiple regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and assume that both $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\mathbf{y}$ are in correlation form. Show that the ridge estimator of $\boldsymbol{\beta}$, denoted by $\widehat{\boldsymbol{\beta}}_{\text{Ridge}}$ can be the obtained as the solution to the constraint optimization problem

   $$\text{Minimize}_{\boldsymbol{\beta}} \left[\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{\text{LS}})' \mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{\text{LS}} \right) \right] \quad \text{subject to} \quad \boldsymbol{\beta}'\boldsymbol{\beta} \leq d,$$

   where $\widehat{\boldsymbol{\beta}}_{\text{LS}}$ is the ordinary least-squares estimator of $\boldsymbol{\beta}$ and $d > 0$ is an arbitrary constant. Sketch the graph (for the two-parameter case) representing the constraint $\boldsymbol{\beta}'\boldsymbol{\beta} \leq d$, explain the role of constant $d > 0$ and the relationship of $\widehat{\boldsymbol{\beta}}_{\text{Ridge}}$ to $\widehat{\boldsymbol{\beta}}_{\text{LS}}$, specifically why $\widehat{\boldsymbol{\beta}}_{\text{Ridge}}$ shrinks the LS estimator $\widehat{\boldsymbol{\beta}}_{\text{LS}}$ towards the origin.

   Hint: Form the function $\phi(\boldsymbol{\beta}) = (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{\text{LS}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{\text{LS}}) + \lambda\boldsymbol{\beta}'\boldsymbol{\beta}$, where $\lambda > 0$ is the Lagrangian multiplier (or ridge parameter). Assuming that $\widehat{\boldsymbol{\beta}}_{\text{LS}}$ is fixed and does not depend on $\boldsymbol{\beta}$, differentiate $\phi(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$, set the result equal to zero and, at the minimum, set $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}_{\text{Ridge}}(\lambda)$.

   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.

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

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

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

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

Variable selection and model building

1. Regression analysis often utilities the variable selection procedure know as the *all possible regressions* (also called for the *best subsets regression*).

   (a) 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 this approach to the regression model building.

   (b) Suppose that there are three candidate predictors, x_1, x_2, and x_3, for the final regression model. Suppose further that the intercept term, β_0 is always included in all the model equations. How many models must be estimated and examined if one applies all possible regressions approach? Motivate you answer.

2. Exercise 10.13 from MPV: (*Hint for part c):* Observe that the correlation for of the variables is used. Recall that for the full model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, with K candidate regressors $x_1, \ldots, x_K$, and with $n \geq K+1$ observations, the following partition can be obtained

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

   where $\mathbf{X}_p$ is an $n \times p$ matrix whose columns represent intercept and $(p-1)$ regressors, $\mathbf{X}_r$ is an $n \times r$ matrix whose columns represent the regressors to be removed from the model, and $\boldsymbol{\beta}_p$ and $\boldsymbol{\beta}_r$ are corresponding parts of $\boldsymbol{\beta}$. Then for the OLS estimator of the coefficients in the reduced model, the following holds

   $$\mathrm{E}(\widehat{\boldsymbol{\beta}}_p) = \boldsymbol{\beta}_p + (\mathbf{X}_p'\mathbf{X}_p)^{-1}\mathbf{X}_p'\mathbf{X}_r\boldsymbol{\beta}_r.$$

   (*Hint for part d):* Recall that the mean square error of an estimate $\hat{\theta}$ of the parameter θ id defined as

   $$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + [\mathrm{E}(\hat{\theta}) - \theta]^2.$$

CART, logistic regression and GLM, bootstrapping in regression

1. Let D_a denote the training data set and let T be a decision tree trained on D_a through an error measuring function E (e.g. SS_{Res} on a regression tree or cross-entropy on a classification tree). Let T' denote the resulting tree after a split on T, and suppose that $E(D_a, T) = E(D_a, T')$. Should the splitting process be stopped? Justify your answer.

2. Let D_a and D_b be two independent data sets. Let T be a decision tree trained on D_a through an error measuring function E (e.g. SS_{Res} on a regression tree or cross-entropy on a classification tree). Describe how to prune a regression or decision tree using information given by $E(D_a, T)$ and $E(D_b, T)$.

3. Consider a continuous (latent) variable Y^* given as follows:

$$Y^* = \boldsymbol{\beta}^{'}\mathbf{x} + \varepsilon$$

where $\boldsymbol{\beta}^{'}\mathbf{x} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$ and $\varepsilon \in N(0,1)$ is independent of $\mathbf{x}$. Define further Y as the indicator

$$Y = \begin{cases} 1 & \text{if } Y^* > 0 \text{ i.e. } -\varepsilon < \boldsymbol{\beta}^{'}\mathbf{x}, \\ 0 & \text{otherwise.} \end{cases}$$

   (a) Show that for all real u,

$$P(-\varepsilon \leq u) = P(\varepsilon \leq u).$$

   (b) Show that

$$P(Y = 1 \mid \mathbf{x}) = \Phi\left(\boldsymbol{\beta}^{'}\mathbf{x}\right),$$

   where $\Phi(\cdot)$ is the distribution function of $N(0,1)$.

   You are likely to need (a) in this. But if you cannot solve (a), you are still allowed to use the formula/result in (a)

4. 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 $\mathbf{x}_i = (1, x_{i1}, x_{i2}, \ldots, x_{ip})$.

   (a) Show that for all real $\boldsymbol{\beta}$, the log likelihood function $l(\boldsymbol{\beta})$ can be written as

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

   (b) Find the partial derivatives $\frac{\partial}{\partial \beta_0} l(\boldsymbol{\beta})$ and $\frac{\partial^2}{\partial \beta_0^2} l(\boldsymbol{\beta})$ in the form they would appear in a recursive algorithm like Newton-Raphson for finding the maximum likelihood estimate.

5. You are a statistics consultant, and has been hired by *Skolverket* to create a model for predicting the yearly number of sick days per student in an upper secondary school class, in different parts of Sweden, based on the *Median Student Age* in the class as well as the *City* where the school is located. *Skolverket* provides you with the data containing k observations, $y_1, \ldots, y_k$, one for each combination of *Median Student Age* and *City*. Each combination is called a "Cell", and y_i denotes the observed number of sick days per student in cell i.

You choose a GLM (generalized linear model) for modeling the mean number of sick days per student. The GLM model has the following form

$$g(\mu_i) = \beta_0 + \sum_{j=1}^{N} x_{ij}\beta_j,$$

where $g(\cdot)$ is the link function, $\mu_i = \mathrm{E}(Y_i)$ is the mean number of sick days for a student qualifying for cell i, β_0 is the base level and N is the number of non-redundant parameters The "dummy variable" x_{ij} takes the value 1 if parameter is included in cell i, else 0.(Obs! These notations are used during the lectures on GLM).

(a) A GLM model requires the response variable to belong to a certain family of distributions. What is the name of this family? Also, specify which sub-family of distributions that the number of sick days during a year is likely to belong to.

(b) What is a suitable choice of the link function? *Hint:* The model is multiplicative.

(c) You choose cell $i = 1$ to be the reference cell. Complete the equation

$$g(\mu_1) = ...$$

(d) Describe the steps of the Maximum Likelihood estimation method for estimating the parameters $\beta_j, j = 0 \ldots, N$ using the observed data $y_1, \ldots, y_k$. Let $f_{Y_i}(y_i)$ denote the probability density function of Y_i.

6. You have just created a GLM model for the mean number of car accidents within an arbitrary geographic area, where you use *Population Density* and *Median age of driver* as rating variables. You have found both variables to be good predictors. Now you consider to add *Average Temperature* as a third variable, categorized in three groups: "Cold", "Medium" and "Hot". Call the first model (excluding temperature) the "Small (reduced) model", RM, and the larger model (including temperature) the "Large (full) model", FM.

   (a) How many more non-redundant parameters β_j does FM include, compared to RM?

   (b) Describe the strategy of Wald inference for testing the significance of the new model parameter(s).

   (c) Explain how you can perform a Likelihood Ratio test to test whether the Large (full) model fits the data significantly better than the Small (reduced) model, using the deviance from each of the models.

   (d) What is the distribution of the Log Likelihood test statistic, used for testing the Large (full) model against the Small (reduced) model in this example?

7. (a) Two main sampling procedures for bootstrapping regression estimates are usually referred to as *bootstrapping residuals* and *bootstrapping cases*. Give in detail 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 how to obtain approximate confidence intervals for regression coefficients though bootstrapping.