UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Examination in: STK3100 / STK4100 — Introduction to generalized linear models.

Day of examination: Wednesday December 18th 2019

Examination hours: 9.00 – 13.00.

This problem set consists of 5 pages.

Appendices: Formulas in STK3100 / STK4100

Permitted aids: Approved calculator

Please make sure that your copy of the problem set is complete before you attempt to answer anything.

Problem 1

a) Assume that $Y_i, i = 1, \ldots, n$ are independent binary responses with $\pi_i = \mathrm{P}(Y_i = 1) = \frac{\exp(\alpha+\beta x_i)}{1+\exp(\alpha+\beta x_i)}$. Here x_i is an explanatory variable and α and β regression parameters.

Show that $\exp(\beta)$ can be interpreted as an odds-ratio.

Give an approximate interpretation of $\exp(\beta)$ valid when the π_i's are small.

b) In a case-control study, where typically the π_i's are small, one will only collect data on the explanatory variable for a subset of the observations but will oversample the observations with $Y_i = 1$. Thus, one collects x_i for observation i if a sampling indicator equals 1, that is $Z_i = 1$, where

$$\rho_1 = \mathrm{P}(Z_i = 1|Y_i = 1) \quad \text{and} \quad \rho_0 = \mathrm{P}(Z_i = 1|Y_i = 0),$$

allowing for different sampling fractions ρ_1 and ρ_0 for cases ($Y_i = 1$) and controls ($Y_i = 0$). Note that ρ_j can not depend on x_i.

Show that

$$\mathrm{P}(Y_i = 1|Z_i = 1) = \frac{\exp(\alpha^* + \beta x_i)}{1 + \exp(\alpha^* + \beta x_i)}$$

where $\alpha^* = \alpha + \log(\rho_1/\rho_0)$.

What does this result mean in practice for analyzing case-control data?

Hint: First show that $P(Z_i = 1) = \rho_1\pi_i + \rho_0(1 - \pi_i)$ or use Bayes theorem directly.

(Continued on page 2.)

Problem 2

a) The density for the gamma distribution can be expressed as $f(y; \mu, k) = (k/\mu)^k y^{k-1} \exp(-(k/\mu)y)/\Gamma(k)$ for $y > 0$.

Show that the gamma distribution density can be rewritten on the exponential dispersion family form $\exp((\theta y - b(\theta))/\phi + c(y, \phi))$ with $\theta = -1/\mu$, $b(\theta) = -\log(-\theta)$ and $\phi = 1/k$.

Verify that when $Y \sim f(y; \mu, k)$ then $\mathrm{E}[Y] = \mu$ and $\mathrm{var}[Y] = \phi\mu^2$.

b) Write down the definition of a generalized linear model with gamma distributed responses $Y_i, i = 1, \ldots, n$.

Verify that the likelihood-equations for such a model can be written as

$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi\mu_i^2} x_{ij} \frac{\partial \mu_i}{\partial \eta_i} = 0 \quad \text{for } j = 1, \ldots, p$$

with observations y_i of Y_i, x_{ij} is explanatory variable $j = 1, \ldots, p$ and η_i the linear predictor for observation i.

c) The estimators of the regression coefficients determined by solving the equations in question b) are valid also when the Y_i's are not gamma distributed as long as the expected values μ_i and the variance structure $\mathrm{var}[Y_i] = \phi\mu_i^2$ are correctly specified.

Give a brief explanation for why this is true.

Suggest an estimator for the dispersion term ϕ which is valid both when the Y_i's are gamma distributed and when only the expectation and variance structure are correctly specified.

d) Prices of $n = 100$ apartments sold in Oslo in the year 2000 were collected along with explanatory variables x_1 = area in m^2 (in `R`-output `size`), x_2 = no. of rooms in the apartment (`rooms`), x_3 = indicator if the apartment has a balcony or not (`balcony`), x_4 = monthly expenses or rent in NOK (`rent`) and x_5 = location of apartment in west/east direction measured in km (low numbers means in west, high in east of Oslo) (`x`). Results from analyzing the prices with a gamma GLM with an identity link are given on the next page.

Give a description of how the explanatory variables affect the price of the apartments.

Identify which explanatory variables significantly influence the price.

Find the estimated apartment price when $x_1 = 70\ \mathrm{m}^2$, $x_2 = 2$ rooms, $x_3 = 0$, i.e. no balcony, $x_4 = 1000$ NOK and $x_5 = 2$ km to the east.

How would you determine the uncertainty of this estimate (an exact numerical answer is not possible with the given information).

(Continued on page 3.)

```
Call:
glm(formula = price ~ size + rooms + balcony + rent + x,
       family = Gamma(link = identity))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 526.64340   48.15544  10.936  < 2e-16 ***
size         18.39589    1.31699  13.968  < 2e-16 ***
rooms        25.77173   30.88332   0.834  0.40612
balcony      81.67536   30.14617   2.709  0.00801 **
rent         -0.12745    0.01368  -9.318 5.19e-15 ***
x           -93.28967   10.29277  -9.064 1.80e-14 ***
---
(Dispersion parameter for Gamma family taken to be 0.0167957)

    Null deviance: 15.5898  on 99  degrees of freedom
Residual deviance:  1.6817  on 94  degrees of freedom
```

e) Since an identity link has been used one may also consider analyzing the apartment prices with linear regression. Below you find results from such an analysis with the same explanatory variables. In addition residual plots from both models are included where the top panels show the deviance residuals and square roots of the absolute values of standardized deviance residuals against predicted values from the gamma fit whereas the bottom panels gives the corresponding plots for the linear regression.

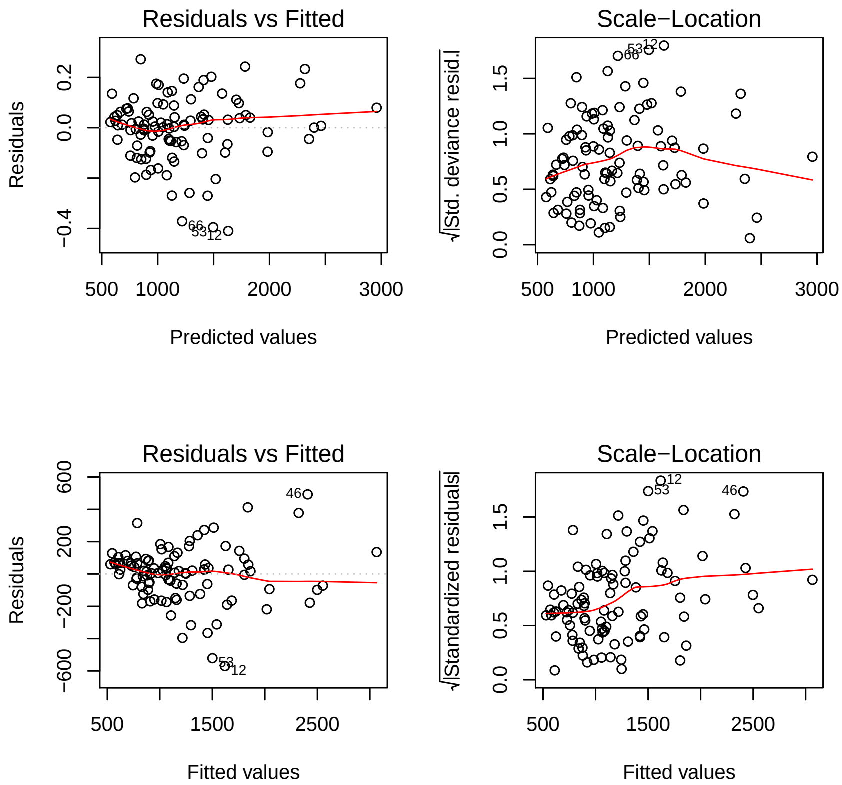
   Give a precise statement of the linear models used here.

   Discuss differences and similarities between the analyses.

```
Call:
lm(formula = price ~ size + rooms + balcony + rent + x)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  525.90793   70.69386   7.439 4.71e-11 ***
size          20.23111    1.37069  14.760  < 2e-16 ***
rooms          3.76506   37.59353   0.100  0.92044
balcony      129.94298   38.97715   3.334  0.00123 **
rent          -0.13969    0.01704  -8.197 1.23e-12 ***
x           -108.17047   13.46564  -8.033 2.72e-12 ***
---
Residual standard error: 174.9 on 94 degrees of freedom
Multiple R-squared:  0.8948,    Adjusted R-squared:  0.8892
F-statistic: 159.9 on 5 and 94 DF,  p-value: < 2.2e-16
```

(Continued on page 4.)



Problem 3

Let Y_{ij} be count response no. $j = 1, \dots, d$ in group $i = 1, \dots, n$. A mixed Poisson model for Y_{ij} with group specific random intercept u_i and one explanatory variable x_{ij} is defined by Y_{ij} being independent and Poisson distributed given u_i with conditional mean

$$\mathrm{E}[Y_{ij}|u_i] = \mu_{ij} = \exp(\beta_0 + \beta_1 x_{ij} + u_i),$$

thus with a log-link for the mixed model and deterministic $\beta_0 + \beta_1 x_{ij}$.

a) Show that marginally

$$\mathrm{E}[Y_{ij}] = \exp(\beta_0 + \beta_1 x_{ij})\mathrm{E}[e^{u_i}]$$

and determine $\mathrm{E}[e^{u_i}]$ when $u_i \sim \mathrm{N}(0, \sigma_u^2)$.

Comment on the relationship between the parameters in the mixed and marginal models.

Hint: The moment generating function of a normal distribution with mean zero and variance σ_u^2 is given by $M(t) = \exp(\sigma_u^2 t^2/2)$.

(Continued on page 5.)

b) Derive an expression for the marginal variance of Y_{ij} under the assumption that $u_i \sim \mathrm{N}(0, \sigma_u^2)$.

c) Assume instead that conditionally on random intercepts $u_i \sim \mathrm{N}(0, \sigma_u^2)$ the responses Y_{ij} are gamma distributed with means $\mathrm{E}[Y_{ij}|u_i] = \exp(\beta_0 + \beta_1 x_{ij} + u_i)$, i.e. still a log-link, and a dispersion term ϕ.

   Derive an expression for the marginal mean $\mathrm{E}[Y_{ij}]$ in this situation.

   Comment on the relationship between the parameters in the mixed and marginal models also in this situation.

   Finally find an expression for the marginal variance of Y_{ij}.

END