Supplementary problems STK3100-f15

Problem S1
A one-way analysis of variance can be formulated as

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \tag{1}$$

where $\sum \alpha_i = 0$ and ϵ_{ij} are i.i.d $N(0, \sigma^2)$. Here Y_{ij} represents the j'th observation from the i'th treatment.

(a) An alternative formulation is

$$Y_{ij} = \mu_i + \epsilon_{ij} \tag{2}$$

Formulate both (1) and (2) using design matrices , i.e. as

$$\underline{Y} = X\underline{\beta} + \underline{\epsilon}.$$

What is the relation between the two parameterizations? Find the (ordinary) least squares estimators in the two cases.

(b) A third possibility is to write

$$Y_{ij} = \kappa + \gamma_i + \epsilon_{ij}$$

where $\gamma_1 = 0$. This is often called "corner-point parameterization". I deJ&H is the category of reference where the factor has level 1 for base level. Formulate also this parameterization using a design matrix. What is the relation with the parameterizations in (1) og (2)? Find the (ordinary) least squares estimators also in this case.

(c) Discuss pros and cons with the three formulations. Are there situations where one alternative is preferable?

Problem S2
A simple linear regression model can be formulated either as

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

or

$$Y_i = a + b(x_i - \bar{x}) + \epsilon_i$$

where $\bar{x} = \frac{1}{N}\sum x_i$ and β_0, β_1, a and b are parameters.

(a) Express both formulations using vectors and matrices, and explain what the design matrices look like. What does the requirement that the design matrix shall be of full rank mean in this situation?

(b) Find the (ordinary) least squares estimators for (β_0, β_1) and (a, b).

(c) Also find the covariance matrices in the two cases, i.e. . $\sigma^2(X'X)^{-1}$ where$X$ is the design matrix. Why is $\mathrm{Cov}(\hat{a}, \hat{b}) = 0$?

Problem S3
A model for a two-way analysis of variance can be written as

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \delta_{ij} + \epsilon_{ijk}$$

where $\sum \alpha_i = 0, \sum \beta_j = 0, \sum_i \delta_{ij} = 0, \sum_j \delta_{ij} = 0$ and ϵ_{ijk} are i.i.d $N(0, \sigma^2)$.
The "corner-point" parameterization is in this case

$$Y_{ijk} = \gamma + \kappa_i + \rho_j + \tau_{ij} + \epsilon_{ijk}$$

where $\kappa_1 = 0, \rho_1 = 0, \tau_{1j} = 0, \tau_{i1} = 0$

(a) Let $I = 3$ and $J = 3$, which correspond to a 3×3 two-way layout. What are the nine linear equations describing the relations between the two parameterizations?

(b) Find the design matrices and explain their relation.

Problem S4
This problem deals with *confounding* which can occur in regression models where one or more co-variates are missing. Assume a linear regression model with two co-variates, $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$.

a) Show that the expectation of $\hat{b}_1$ in a simple linear regression $y = b_0 + b_1 x_1 + \epsilon$ has expectation

$$b_1 = \beta_1 + \beta_2 \rho \frac{\sigma_2}{\sigma_1}$$

where $\sigma_1^2 = \sum_{i=1}^n (x_{i1} - \bar{x}_1)^2/n$, $\sigma_2^2 = \sum_{i=1}^n (x_{i2} - \bar{x}_2)^2/n$,
$\rho = \frac{\sum_{i=1}^n (x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2)}{\sqrt{\sum_{i=1}^n (x_{i1}-\bar{x}_1)^2 \sum_{i=1}^n (x_{i1}-\bar{x}_1)^2}}$ so ρ is the correlation between x_1 and x_2 and σ_j the average sum of squares around the mean for the co-variates x_j.

b) Suppose X_1 and X_3 are independent standard normal random variables. Let $X_2 = \rho X_1 + \sqrt{1-\rho^2} X_3$. Show that $\mathrm{E}(X_2) = 0, \mathrm{Var}(X_2) = 1$ and $\mathrm{Cov}(X_1, X_2) = \rho$. Simulate 100 pairs of co-variates (x_{i1}, x_{i2}) from the simultaneous distribution of (X_1, X_2) with a correlation equal to $\rho = 0.5$. Compute the empirical (Pearson) correlation r between x_{i1} and x_{i2}.

c) From these 100 pairs of co-variates simulate 100 responses $Y_i = x_{1i} + x_{2i} + \epsilon_i$ where $\epsilon_i \sim \mathrm{N}(0, 1)$. Run a simple linear regression against x_{1i}. Test the null hypotheses $b_1 = 1$ and $b_1 = 1.5$. Explain why this is the result you should expect.

d) Consider now the simultaneous regression using both co-variates. Verify that

$$\hat{b}_1 = \hat{\beta}_1 + r \frac{s_2}{s_1} \hat{\beta}_2$$

where r s_j is empirical standard deviation for x_{ij}-ene and $\hat{\beta}_j$ is the ordinary least squares estimators in the regression using both co-variates.

e) Suppose now x_{i2} only has an effect on Y_i, but it is the regression model $Y_i = b_0 + b_1 x_{i1}$ that is analyzed. Which effect are estimated? Explain why it is wrong to identify a significant effect with a causal effect.

Problem S5
Consider a situation where the dispersion parameters for unit i can be written $\phi_i = a_i\phi$ where ϕ is unknown and $a_1, \cdots, a_n$ are known constants. We shall consider how the maximum likelihood estimates for ϕ can be found.

a) Explain why the log likelihood have the form

$$l(\beta_0, \cdots, \beta_p, \phi) = \phi^{-1} \sum_{i=1}^{n} [\frac{y_i\theta_i - a(\theta_i)}{a_i} + c(y_i, a_i\phi)].$$

b) Assume ϕ and $\beta_0, \cdots, \beta_p$ are functionally independent, i.e. does not depend on each other. Show that the likelihood equations for $\beta_0, \cdots, \beta_p$ are independent of ϕ. Explain why the profile likelihood for ϕ, $l_P(\phi) = \max_{\beta_0,\cdots,\beta_p} l(\beta_0, \cdots, \beta_p, \phi)$ is

$$l_P(\phi) = \phi^{-1} \sum_{i=1}^{n} [\frac{y_i\hat{\theta}_i - a(\hat{\theta}_i)}{a_i} + c(y_i, a_i\phi)].$$

c) Show that for gamma data the maximum likelihood estimator of ν solves $\log\nu - \psi(\nu) = \sum_{i=1}^{n}(z_i - \log z_i - 1)/n$ where $z_i = y_i/\hat{\mu}_i$ and $\psi(\nu)$ is the digamma function $\partial \log\Gamma(\nu)/\partial\nu$

Problem S6
Show that in a generalized model with unknown dispersion parameter the elements $E[-\partial^2 l/\partial\beta_j\partial\phi] = 0$ in the Fisher information matrix. What is the implication for the maximum likelihood estimators $\hat{\phi}$ and $\hat{\beta}_j$?

Problem S7
Explain why the expected and observed information matrix are the same in a GLM model with canonical link and known dispersion parameter. The observed information matrix is -Hessian, where Hessian is the matrix of cross derivatives of the log likelihood.

Problem S8
Fit a model with Poisson response, canonical link, ie. log link and with linear

predictor

$$\eta_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2,$$

where x_i is age, to the data in the file `Birth` på available on the course web page.

a) Explain how the null hypotheses $H_0 : \beta_2 = 0$ against $H_1 : \beta_2 \neq 0$ can be tested with a likelihood ratio and a Wald test. What are the P-values for the tests?

b) Do the same as in part a) but now for the null hypotheses

$$\mathrm{H}_0 : \beta_1 = \beta_2 = 0.$$

(Hint: For the Wald test you need the covariance matrix of $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)'$ which you get from the command `vcov(modx)` when `modx` is the fitted model. To fit a quadratic function you may use the formula `children`$\sim$`age +I(age`^{2}`)` in the procedure `glm`)

Problem S9

Let Y_1, Y_2 and Y_3 be independent Poisson distributed variables with expectations μ_1, μ_2 and μ_3.

a) Find the conditional distribution of Y_1 given $Y_1 + Y_2 = y$.

b) What is the conditional distribution of (Y_1, Y_2, Y_3) given $Y_1 + Y_2 + Y_3 = y$?

Problem S10

Let X and Y have simultaneous density/frequency function $f_{x,y}(x, y)$ and assume that $E|Y| < \infty$.

a) Show the rule for iterated expectation: $E[Y] = E[E[Y|X]]$.

b) Show then $Var(Y) = Var(E[Y|X]) + E[Var(Y|X)]$.

Problem S11
Use the results from the previous problem to find the expectation and variance of the negative binomial distributed variable defined in deJong and Heller on page 24 and 31-32. What is the expectation and variance of a Poisson-inverse Gaussisian distributed variable defined in de Jong & Heller on page 33?

Problem S12

Consider a situation where $y_1, \ldots, y_n$ are independently, identically $N(\mu, \sigma^2)$ distributed. The parameters μ and σ^2 are unknown. Let $z_i = y_i - \bar{y}$, $1 = 1, \ldots, n-1$ where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. We shall see how an estimator for σ^2 can be constructed from $z_i, i = 1, \ldots, n-1$.

a) Show that the covariance matrix of $z = (z_1, \ldots, z_{n-1})'$ can be expressed as
$$\Sigma_z = \sigma^2(I_{n-1} - 1_{n-1}1'_{n-1})$$
where I_{n-1} is the $(n-1) \times (n-1)$ identity matrix and 1_{n-1} is an $n-1$ dimensional vector with elements equal to 1.

b) Verify that
$$\Sigma_z^{-1} = \frac{1}{\sigma^2}(I_{n-1} + 1_{n-1}1'_{n-1}).$$

c) Show that the likelihood for $z_1, \ldots, z_{n-1}$ can be written
$$L_z(\sigma^2) = \frac{1}{(2\pi)^{(n-1)/2}\sigma^{(n-1)}} \exp(-\frac{1}{2\sigma^2} z'(I_{n-1} + 1_{n-1}1'_{n-1})z),$$
and that the maximum likelihood estimator of σ^2 based on $z_1, \ldots, z_{n-1}$ is equal to
$$\hat{\sigma}^2 = \frac{1}{n-1} z'(I_{n-1} + 1_{n-1}1'_{n-1})z.$$

d) Show that $\sum_{i=1}^{n-1} z_i = -(y_n - \bar{y})$ so that $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$.

e) Discuss how these results relate to REML estimation?

Problem S13

Let $(y_i, x_{i1}, \ldots, x_{ip}),\ i = 1, \ldots$ be observations which are modeled as a GLM model where the responses y_i are gamma distributed with density

$$f(y; \mu, \nu) = \frac{1}{\Gamma(\nu)} y^{\nu-1} (\frac{\nu}{\mu})^{\nu} \exp(-\frac{\nu y}{\mu}), \quad y, \mu, \nu > 0.$$

The predictor is $\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$.

a) Show that the maximum likelihood estimators for $\beta_0, \beta_1, \ldots, \beta_p$ and ν satisfy

$$2n[\log(\hat{\nu}) - \psi(\hat{\nu})] = \Delta/\hat{\nu}$$

where Δ is the deviance as defined in deJong and Heller and $\psi(\nu)$ is the di-gamma function $\Gamma(\nu)'/\Gamma(\nu)$.

b) Show that the asymptotic distributions of the maximum likelihood estimators of ν and $\beta_0, \beta_1, \ldots, \beta_p$ are independent.

Problem S14, final STK3100-f10, problem 2

The table displayed below is from a famous data set from an investigation of the relation between smoking and heart diseases among British physicians. The number of deaths is the response and age (`age`) and smoking (`smoker`) are the covariates. Let age be a numerical variable and associate the values, or scores, 40, 50, 60, 70 and 80 to the five age groups. Smoking is a factor with two levels where 0 designates non-smoker and 1 smoker. In addition a variable (`persyear`) which indicates the number of years lived for the different categories is recorded.

Table 1: Table 1: Mortality and smoking.

	Person year		Heart related death	
Age	Non-smoker	Smoker	Non-smoker	Smoker
35-44	18793	52407	2	32
45-54	10673	43248	12	104
55-64	5710	28612	28	206
65-74	2585	12633	28	186
75-84	1462	5317	31	102

The output displays the result from fitting a model where the response is Poisson distributed and the canonical link is used.

```
Call:
glm(formula = deaths ~ offset(log(persyear)) + I(age) + I(age^2) +
factor(smoker) + I(age):factor(smoker), family = poisson, data = coro)

Coefficients:

                      Estimate Std. Error   z value  Pr(>|z|)
(Intercept)            -1.971e+01 1.253e+00 -15.734  < 2e-16 ***
I(age)                  3.565e-01 3.631e-02   9.819  < 2e-16 ***
I(age^2)               -1.978e-03 2.736e-04  -7.228  4.89e-13 ***
factor(smoker)1         2.370e+00 6.559e-01   3.613  0.000303 ***
I(age):factor(smoker)1 -3.084e-02 9.699e-03  -3.180  0.001474 **
---

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 936.6589 on 9 degrees of freedom
Residual deviance: 1.6661 on 5 degrees of freedom
AIC: 66.734
```

a) Explain why the assumption that the response is Poisson distributed is reasonable in this situation. Give an explicit description of the role of

the covariates in the model that is fitted in the output above. Comment on the results.

b) Explain what `offset` is. Why is it sensible to use `offset` in this case?

c) Express the effect of smoking for this kind of mortality by the relevant rate ratios. Compute especially the ratios between smokers and non-smokers for physicians that are 40 years old and for physicians that are 70 years old. an Discuss the result.

Below is the output from fitting a more general model.

```
Call:
glm(formula = deaths ~ offset(log(persyear)) + I(age) + I(age^2) +
factor(smoker) + I(age):factor(smoker) + I(age^2):factor(smoker),
family = poisson, data = coro)

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -2.153e+01 3.197e+00 -6.736  1.63e-11 ***
I(age)                    4.148e-01 1.004e-01  4.130  3.62e-05 ***
I(age^2)                 -2.430e-03 7.739e-04 -3.140  0.00169 **
factor(smoker)1           4.445e+00 3.391e+00  1.311  0.18991
I(age):factor(smoker)1   -9.755e-02 1.069e-01 -0.912  0.36160
I(age^2):factor(smoker)1 5.196e-04 8.273e-04  0.628  0.52999
---

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
(Dispersion parameter for poisson family taken to be 1)

Null deviance: 936.6589 on 9 degrees of freedom
Residual deviance: 1.2623 on 4 degrees of freedom

AIC: 68.33
Number of Fisher Scoring iterations: 4
```

d) The two last to estimates that describes interaction between age and smoking are not individually significant. Do a Wald test to test the

hypotheses that the corresponding coefficients are both equal to zero, ie. the simultaneous hypotheses that both are equal to zero. The matrix below is the estimated covariance matrix for the estimators of the two coefficients.

```
                            I(age):factor(smoker)1 I(age^2):factor(smoker)1
I(age):factor(smoker)1    1.143363e-02           -8.807653e-05
I(age^2):factor(smoker)1 -8.807653e-05            6.844424e-07
```

The inverse covariance matrix has diagonal elements 10038.02 and 16768.5354e+04. The off-diagonal elements are both 12917.2852e+02.

e) Explain why the expected and observed information matrices are equal in models treated in this problem.

Problem S15, final STK3100-f09, problem 2

In a German survey people were asked about the number of visits to physicians during the last three month (`numvisit`) and about various covariates that may affect this number for 1100 women. Among these covariates we shall only consider an indicator for self reported bad health (`badh`) and age (`age`) in years.

(a) Below are the results from a fit for a Poisson model using `R` with a log link, where the response is `numvisit` and the covariates `badh` and age `age`

Describe the model formally and give an interpretation of the parameters.

```
> M0<-glm(numvisit~age+badh,family=poisson)
> summary(M0)
Deviance Residuals:

Min 1Q Median 3Q Max
-3.9452 -2.0348 -0.8169 0.5191 12.5571
```

```
Coefficients:

             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.588731 0.064318   9.153  < 2e-16 ***
age          0.005556 0.001676   3.316  0.000914 ***
badh         1.140908 0.039858  28.625  < 2e-16 ***
---
Signif. codes: 0 ?e*** ?? 0.001 ?e** ?? 0.01 ?e* ?? 0.05 ?e. ?? 0.1 ?e ??
(Dispersion parameter for poisson family taken to be 1)

Null deviance: 4779.4 on 1099 degrees of freedom
Residual deviance: 3975.3 on 1097 degrees of freedom
```

(b) Use the output from the R-fit to find estimates for rate ratios and 95% confidence intervals for visit to physicians for
(i) women with self reported bad health and self reported good health.
(ii) women of age 50 and women of age 40 years.
Estimate also the rate for for visit to physicians for a woman who is 40 years old and reports to be in good health. What additional information do you need to find a confidence interval for the (theoretical) rate?

(c) A more general model allows for overdispersion compared to the Poisson model in part a) via a specification of $Var(Y) = \phi\mu$ where Y is the response, the number of visits to physicians and μ is the expectation of Y A method to take overdispersion into account is to assume that $Y|Z$ is Poisson distributed with expectation $Z\mu$ where Z is a latent variable. What is the marginal distribution if Z is gamma distributed with expectation 1? What are the reasons for your answer?

Problem S16, final STK3100-f08

(a) Show that the frequency function of a Poisson distributed variable can be written as $f(y;\theta) = c(y)exp(\theta y - a(\theta))$. What is the relation between

the expectation μ of a Poisson distributed variable and the parameter θ. Find explicit expressions for the functions $a(\theta)$ and $c(y)$.

(b) Suppose that $Y_1, \ldots, Y_n$ is independent Poisson distributed variables with expectation $\mu_i = \exp(\alpha + \beta x_i$ where α and β are regression parameters and x_i are known covariates. Show that this is a particular case of generalized models (GLM).

(c) What is the log-likelihood for the data in part (b)? Show that the score function can be written as

$$U(\alpha, \beta) = \begin{pmatrix} U_1(\alpha, \beta) \\ U_2(\alpha, \beta) \end{pmatrix} = \sum_{i=1}^{n} \begin{pmatrix} 1 \\ x_i \end{pmatrix} (Y_i - \mu_i).$$

Find also an expression for the expected information matrix.

(d) Explain what an saturated model is. Show that the maximum likelihood estimators for $\mu_i, i = 1, \ldots, n$ are $\tilde{\mu}_i = Y_i$ in the saturated model

(e) Use this result to find an expression for the deviance in a generalized linear model with responses which are Poisson distributed. Explain what kind of tests that can be performed using deviances.

(f) Consider a situation where the focus is not on the number of accidents, but the analysis is based on variables indicating the occurrence of at least one accident, i.e. $Y_i' = I(Y_i > 0)$. Let $\pi_i = P(Y_i' = 1)$. Show that this is a generalized linear model as in part (b) with link function

$$g(\pi_i) = \log(-\log(1 - \pi_i)).$$

What is this link function called?

(g) Below is the results from an analysis of an vehicle insurance policy where the response is whether the insured has reported one or more accidents last year. The model for the observations is a Poisson regression model as described in part (b), but in the analysis the binary responses corresponding to the model developed in part (f) are used. The covariates are the age of driver divided in 6 groups (`agecat`), and the value of the car in units of 10.000 dollars (`veh-value`) (used as a

numerical variable so that a car worth 25.000 dollars is coded as 2.5). Only cars worth less than 40.000 dollars are included in the analysis.

Use the R-output to compute the relative difference in the rates of accidents between

(i) age category 2 and age category 1

(ii) age category 5 and age category 2

(iii) two cars worth 25.000 and 5.000 dollar respectively

(iv) two drivers where one is in age category 2 and owns a car worth 20.000 dollar, while the other driver belongs to age category 1 and owns a car worth 10.000 dollars

```
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.79067 0.05636 -31.773 < 2e-16 ***
factor(agecat)2 -0.21697 0.05846 -3.711 0.000206 ***
factor(agecat)3 -0.25327 0.05674 -4.464 8.06e-06 ***
factor(agecat)4 -0.27294 0.05663 -4.820 1.44e-06 ***
factor(agecat)5 -0.51762 0.06396 -8.093 5.83e-16 ***
factor(agecat)6 -0.47167 0.07183 -6.567 5.14e-11 ***
veh-value 0.11656 0.01874 6.220 4.96e-10 ***
```

h) Find also a 95% confidence interval for the relative change in the rate for accidents between

(i) age category 2 and alders category 1

(iv) two drivers where one driver belongs to age category 2 and owns a car worth 20.000 dollars, while the other driver belongs to age category 1 and owns a car worth 10.000 dollars

For (iv) you need the coefficient of correlation between the estimated coefficients corresponding to age group 2 and the value of the car is -0.0259.

Problem S17, final STK3100-f10, problem 1

In this problem we shall consider a record from 35 operations where the occurrence of sore throat after narcosis is recorded. Sore throat, `sore`, is response, and is coded as 1 if it occurred and as 0 otherwise. The covariates are the duration (`duration` in minutes of the operation and two types of equipment (`type`) used to keep the throat open during the operation.

The output below is based on a model assuming that the binominally distributed, $Bin(m, \pi)$, where m is equal to 1, i.e. binary responses.

The link function is logit link. We begin by only considering duration as covariate, i.e. with a predictor of the form $\eta = \beta_0 + \beta_1 x$ where x is the duration of the operation.

```
Call:
glm(formula = sore ~ I(duration), family = binomial, data = s
ore)
Deviance Residuals:
Min 1Q Median 3Q Max
-2.0964 -0.7392 0.3020 0.8711 1.3753
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.21358 0.99874 -2.216 0.02667 *
I(duration) 0.07038 0.02667 2.639 0.00831 **
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 46.180 on 34 degrees of freedom
Residual deviance: 33.651 on 33 degrees of freedom
AIC: 37.651
```

a) Explain how a generalized linear model is defined and why the model described above is of this type.

b) Estimate the odd ratio between two operations, where one has duration 30 minutes and the other 40 minutes. Also find a 95% confidence

interval for this odds ratio.

c) What is the predicted probability of sore throat in an operation of duration 40 minutes? Find also a 95% confidence interval. Here you need that the estimated correlation between $\hat{\beta}_0$ and $\hat{\beta}_1$ is -0.906.

d) In the analysis of deviance analysis table below you find the deviance for models also containing the covariate `type` and a quadratic term in duration. The entries for the degrees of freedom are erased. Fill out what is missing. Then explain that the model considered in parts a)-c) is reasonable. You can assume that the most general model (Model 4) is fitting well.

```
Analysis of Deviance Table
Model 1: sore ~ 1
Model 2: sore ~ I(duration)
Model 3: sore ~ I(duration) + factor(type)
Model 4: sore ~ I(duration) + I(duration^2) + factor(type)
Resid. Df Resid. Dev Df Deviance
1 ? 46.180
2 ? 33.651 ? 12.528
3 ? 30.138 ? 3.513
4 ? 30.133 ? 0.005
```

e) Let y_i and $\hat{\pi}_i$, $i = 1, \ldots, 35$ be the observed and fitted probabilities respectively. Show that the deviance for binomial models with binary response can be expressed as

$$-2\sum_{i=1}^{35}[\hat{\pi}_i \log \frac{\hat{\pi}_i}{1-\hat{\pi}_i} + \log(1-\hat{\pi}_i)].$$

Explain carefully why the deviance is not a useful measure for goodness-of-fit in this case.

Problem S18, final STK3100-f09, problem 1

(a) Show that the binomial distribution (n, π) belongs to the family of exponential distributions where the density can be expressed as

$$f(y; \theta, \phi) = c(y, \phi) \exp((y\theta - a(\theta))/\phi)$$

Find $a(\theta)$ and ϕ and use this to verify the known formula for expectation and variance in the binomial distribution.

(b) Define the concept of a generalized linear model (GLM). What is meant by canonica1 link? Find the canonical link in a binary regression model, i.e. a model where the response has a binary distribution/ Bernoulli distribution. What other link functions are common in such distributions?

(c) Define the concepts saturated model and deviance. Consider a logistic regression model with regression part $\eta_i = \alpha + \beta x_i$, binary response variable y_i and covariates x_i, $i = \ldots, n$. How would you test the hypothesis $H_0 : \beta = 0$ in this model using the deviance concept?

(d) Set up the log likelihood function for the model in part c). Show that the score function with respect to the parameters α and β has components

$$\sum_{i=1}^{n}(y_i - \frac{\exp^{\eta_i}}{1 + \exp^{\eta_i}}) \text{ and } \sum_{i=1}^{n}(x_i y_i - \frac{x_i \exp^{\eta_i}}{1 + \exp^{\eta_i}}).$$

Find the Fisher information matrix. Explain how you can use these quantities to define an iterative procedure to determine the maximum likelihood estimates of the vector $(\alpha, \beta)^T$.

Problem S19, final STK3100-f07, problem 1

A class of exponential distributions can be parameterized with density/ frequency functions on the form

$$f(y;\theta,\phi) = \exp(\theta y - c(\theta) + d(y))$$

(a) Show that the Poisson distributions can be written on this form.

(b) Show that also the exponential distributions can be written on this form.

(c) Find the expectation and variance of the exponential distribution by help of the characterization in part (b). Describe how the variance depends on the expectation.

(d) A Pareto distributed variable has cumulative distribution function $F(y;\lambda) = 1 - (1/y)^{\lambda}$ when $y > 1$. Show that $V = \log(Y)$ is exponentially distributed with expectation $\mu = 1/\lambda$.

(e) Suppose that $Y_i,\ i = 1,\ldots,n$ are independent and Pareto distributed with parameter $\lambda_i = \exp(\alpha + \beta x)$ with known covariates x_i. Explain how one can use a statistical package with the usual implementation of generalized linear models (GLM), such as `R` for example, to estimate the parameters α and β.

(f) More generally one can define a class of exponential distributions with densities of the form

$$g(y;\theta) = \exp(\theta(y) - c(\theta) + d(y)).$$

Show that with this definition the Pareto distributions belong to the exponential class of distributions.

(g) Suppose Y has a density of the form $g(y;\theta) = \exp(\theta a(y) - c(\theta) + d(y))$. Find an expression for the density of $V = a(Y)$. Explain in particular the relation with the parameterization $f(v;\theta)$ from the start of this problem.

Problem S20, final STK3100-f07, problem 2

In this problem we will consider the mortality risk in the so-called post neo-natal period, from the 28. day of life until the first birthday. We shall only consider the sudden infant death syndrome, SIDS, and we model the probability for a SIDS death in the period, given that the child does not die from another cause, using logistic regression.

We will consider how the SIDS-deaths are related to year of birth, gender and birth weight. Year of birth (`kohort`) is recorded as a factor with 5 levels where level 1 corresponds to 1967-1974, level 2 to 1975-1979, level 3 to 1980-1984, level 4 top 1985-1989 and level 5 to 1990-1995. Gender (`kjønn`) is coded as 1 for boys and 2 for girls. Weight of birth (`vekt`) are used as a continuous covariate measured in kilo.

a) Under is the deviance table from a R-output where some quantities are marked with "?". Fill out the missing values and give an interpretation of the results. For p-values it is sufficient to indicate whether they are significant or not. Explain why testing of significance can be performed using the deviance table.

```
Analysis of Deviance Table
Model: binomial, link: logit
                          Df Deviance Resid. Df Resid. Dev    P(>|Chi|)
NULL                                        570     1101.92
vekt                       ?        ?       569      842.33          ?
factor(kohort)             ?        ?         ?      527.74          ?
kjonn                      ?        ?         ?      434.93          ?
vekt:factor(kohort)        ?        ?         ?      428.56          ?
vekt:kjonn                 ?        ?         ?      428.37          ?
factor(kohort):kjonn       ?        ?         ?      413.05     0.0041
vekt:factor(kohort):kjonn  ?        ?         ?      407.80          ?
```

b) Below is the R-output where only the main effects of the covariates birth weight, gender and year of birth is included. For year of birth a corner point parameterization is used where the level 1967-1974 is used as a reference. Give an interpretation of the results using odd ratios computed from the table. Also find a 95% confidence interval for the odds ratio corresponding to birth weight.

```
Coefficients:
                  Estimate     Std. Error  z value    Pr(>|z|)
(Intercept)       -4.37607       0.15833   -27.639    < 2e-16 ***
vekt              -0.67110       0.03758   -17.859    < 2e-16 ***
kjonn             -0.47371       0.04981    -9.511    < 2e-16 ***
factor(kohort)2    0.56224       0.08629     6.515      7.25e-11 ***
factor(kohort)3    0.90941       0.08105    11.220     < 2e-16 ***
factor(kohort)4    1.07958       0.07743    13.943     < 2e-16 ***
factor(kohort)5    0.11049       0.08958     1.233      0.217
```

c) So far we have only considered SIDS-deaths. Suppose there are J different causes of death, and we use a multinominal regression model with J+1 categories, also including children that survive and π_{ij}, $j = 1, \ldots, J$ as the probability that individual i dies of the j'th cause and π_{i0} as the probability that the i'th individual survives. The dependence of the vector of covariates x_i can be expressed as

$$\begin{aligned} \pi_{ij} &= \frac{\exp(\beta_j x_i)}{1 + \sum_{k=1}^{J} \exp(\beta_k x_i)}, \; j = 1, \ldots, J \\ \pi_{i0} &= \frac{1}{1 + \sum_{k=1}^{J} \exp(\beta_k x_i)}, \end{aligned}$$

where β_j is the vector of regression parameters. Let SIDS correspond to $j = 1$.

Show that we have an ordinary logistic regression model if we (as in parts a) and b)) only consider those that die of SIDS and those who survive. Explain what the parameters in the logistic regression model look like.

Discuss drawbacks and advantages by analyzing the data using logistic regression instead of using the full multinominal model