Additional exercise 1

Let $\mathbf{a} = (a_1, a_2, \ldots, a_p)^{\mathrm{T}}$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^{\mathrm{T}}$ be $p \times 1$ vectors, and let $\mathbf{A} = \{a_{ij}\}$ be a $p \times p$ matrix. Then $\partial(\mathbf{a}^{\mathrm{T}}\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ is the $p \times 1$ vector with j-th element $\partial(\mathbf{a}^{\mathrm{T}}\boldsymbol{\beta})/\partial\beta_j$ and $\partial(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{A}\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ is the $p \times 1$ vector with j-th element $\partial(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{A}\boldsymbol{\beta})/\partial\beta_j$. Show that $\partial(\mathbf{a}^{\mathrm{T}}\boldsymbol{\beta})/\partial\boldsymbol{\beta} = \mathbf{a}$ and $\partial(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{A}\boldsymbol{\beta})/\partial\boldsymbol{\beta} = (\mathbf{A}+\mathbf{A}^{\mathrm{T}})\boldsymbol{\beta}$.

Additional exercise 2

Assume $\mathbf{Y} = (Y_1, \ldots, Y_p)^{\mathrm{T}} \sim N(\boldsymbol{\mu}, \boldsymbol{V})$, where $\boldsymbol{V}$ is a positive definite $p \times p$ matrix.

a) Prove that $(\mathbf{Y} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{V}^{-1}(\mathbf{Y} - \boldsymbol{\mu}) \sim \chi^2_p$

b) Prove that $\mathbf{Y}^{\mathrm{T}}\boldsymbol{V}^{-1}\mathbf{Y} \sim \chi^2_{p,\lambda}$, where $\lambda = \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{\mu}$

Hint: Consider the transformation $\mathbf{Z} = \boldsymbol{V}^{-1/2}(\mathbf{Y} - \boldsymbol{\mu})$ in question a and the transformation $\mathbf{Z} = \boldsymbol{V}^{-1/2}\mathbf{Y}$ in question b. Here $\boldsymbol{V}^{1/2}$ is a $p \times p$ symmetric matrix such that $\boldsymbol{V}^{1/2}\boldsymbol{V}^{1/2} = \boldsymbol{V}$ and $\boldsymbol{V}^{-1/2} = (\boldsymbol{V}^{1/2})^{-1}$.

Additional exercise 3

We assume that $Y_1, \ldots, Y_n$ are independent and normally distributed with common variance σ^2 and

$$\mu_i = E(Y_i) = \beta_0 + \beta_1(x_i - \bar{x}),$$

where $\bar{x} = n^{-1}\sum_{i=1}^n x_i$ and not all the x_i are equal.

a) Show that the model matrix may be written $\boldsymbol{X} = [\mathbf{1}_n\,,\, \boldsymbol{x} - \bar{x}\mathbf{1}_n]$, where $\boldsymbol{x} = (x_1, \ldots, x_n)^{\mathrm{T}}$ and $\mathbf{1}_n$ is a n-dimensional vector of 1's.

   What is the rank of the model matrix?

b) Show that the projection matrix onto the model space $C(\mathbf{X})$ may be given as

$$\boldsymbol{P}_{\boldsymbol{X}} = n^{-1}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}} + M^{-1}(\boldsymbol{x} - \bar{x}\mathbf{1}_n)(\boldsymbol{x} - \bar{x}\mathbf{1}_n)^{\mathrm{T}},$$

   where $M = \sum_{i=1}^n (x_i - \bar{x})^2$.

c) Determine the fitted values $\widehat{\boldsymbol{\mu}} = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{Y}$.

d) Let $\boldsymbol{P}_0 = n^{-1}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}$ be the projection matrix for the null model (cf. section 2.3.1 in the book by Agresti). From the orthogonal decomposition

$$\boldsymbol{Y} = \boldsymbol{P}_0\boldsymbol{Y} + (\boldsymbol{P}_{\boldsymbol{X}} - \boldsymbol{P}_0)\boldsymbol{Y} + (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Y}$$

   we obtain

$$\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y} = \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{P}_0\boldsymbol{Y} + \boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{P}_{\boldsymbol{X}} - \boldsymbol{P}_0)\boldsymbol{Y} + \boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Y}$$

   Express the terms in this sum of squares decomposition by $\boldsymbol{Y}$ and $\widehat{\boldsymbol{\mu}}$.

e) Use Cochran's theorem to show that

$$\boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{P_X} - \boldsymbol{P}_0)\boldsymbol{Y}/\sigma^2 \quad \text{and} \quad \boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P_X})\boldsymbol{Y}/\sigma^2$$

are independent and determine their distributions.

f) Derive a F-statistic for testing the null hypothesis $H_0 : \beta_1 = 0$ versus the alternative $H_A : \beta_1 \neq 0$, and determine the distribution of the statistic under H_0 and under H_A.

Additional exercise 4

Consider the normal linear model, where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T \sim N(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$ and the mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$ is given by $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$ for a model matrix $\boldsymbol{X}$ and a vector of parameters $\boldsymbol{\beta}$. Assume that we have observed $\boldsymbol{Y} = \boldsymbol{y} = (y_1, \ldots, y_n)^T$.

a) Show that the likelihood is given by

$$\ell(\boldsymbol{\mu}, \sigma) = (2\pi)^{-n/2}\sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{\mu})^T(\boldsymbol{y} - \boldsymbol{\mu})\right\}. \tag{1}$$

b) Explain that, for any value of σ, the value of $\boldsymbol{\mu}$ that maximizes the likelihood (1) is the vector of fitted values $\widehat{\boldsymbol{\mu}}$ obtained by least squares.

c) When we insert $\widehat{\boldsymbol{\mu}}$ for $\boldsymbol{\mu}$ in (1), we obtain the profile likelihood for σ:

$$\ell_p(\sigma) = \ell(\widehat{\boldsymbol{\mu}}, \sigma) = (2\pi)^{-n/2}\sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y} - \widehat{\boldsymbol{\mu}})^T(\boldsymbol{y} - \widehat{\boldsymbol{\mu}})\right\}.$$

Show that the profile likelihood is maximized for $\widehat{\sigma}^2 = \mathrm{SSE}/n$, where

$$\mathrm{SSE} = (\boldsymbol{y} - \widehat{\boldsymbol{\mu}})^T(\boldsymbol{y} - \widehat{\boldsymbol{\mu}}) = \sum_{i=1}^{n}(y_i - \widehat{\mu}_i)^2.$$

d) Show that the maximum value of the likelihood (1) is given by

$$\ell(\widehat{\boldsymbol{\mu}}, \widehat{\sigma}) = (2\pi)^{-n/2}\widehat{\sigma}^{-n}e^{-n/2}.$$

We now consider two models, denoted M_0 and M_1, with model matrices $\boldsymbol{X}_0$ and $\boldsymbol{X}_1$ of ranks p_0 and p_1, respectively. We assume that the models are nested, so that model M_0 is a special case of model M_1. (More precisely, we have that the model space $C(\boldsymbol{X}_0)$ of model M_0 is a subspace of the model space $C(\boldsymbol{X}_1)$ of model M_1.) We want to test the null hypothesis that model M_0 holds versus the alternative that model M_1 holds (so model M_1 is assumed to hold a priori).

The likelihood ratio test rejects the null hypotesis for small values of the statistic

$$\Lambda = \frac{\max_{M_0} \ell(\boldsymbol{\mu}, \sigma)}{\max_{M_1} \ell(\boldsymbol{\mu}, \sigma)} = \frac{\ell(\widehat{\boldsymbol{\mu}}_0, \widehat{\sigma}_0)}{\ell(\widehat{\boldsymbol{\mu}}_1, \widehat{\sigma}_1)}.$$

Here $\widehat{\boldsymbol{\mu}}_m$ and $\widehat{\sigma}_m$ are the maximum likelihood estimates under model M_m for $m = 0, 1$.

e) Show that

$$\Lambda = \left(1 + \frac{p_1 - p_0}{n - p_1} F\right)^{-n/2}, \tag{2}$$

where F is given by formula (3.1) in the book by Agresti.

Relation (2) shows that Λ is a strictly decreasing function of F. Thus rejecting the null hypothesis for small values of the likelihood ratio statistic Λ corresponds to rejection for large values of the F statistic.

Additional exercise 5
In class we discussed the model

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_i$$

for the beetle data. Here x_i is dose in group i ($i = 1, 2, \ldots, 8$) and π_i is the probability that a beetle in group i will die. The data are available at the homepage of the book, and they may be read into `R` by the command

```
data="http://www.stat.ufl.edu/~aa/glm/data/Beetles2.dat"
beetle=read.table(data,header=T)
```

a) Try out the alternative model

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2.$$

Plot the fitted curve together with the curve from the initial linear model. Check if x_i^2 is significant by looking the related P-value.

One may show that $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2)^{\mathrm{T}}$ is approximately multivariate normally distributed with a covariance matrix that may be estimated by the inverse information matrix. In `R` you may obtain the estimated covariance matrix by the command `summary(fit.a)$cov.scaled` when `fit.a` is the fitted model in question a).

b) Find the (estimated) correlations between the different β-estimates. Why are the correlations that strong?

c) An alternative to the logit link is the probit link given by

$$\Phi^{-1}(\pi_i) = \eta_i,$$

where Φ is the cumulative distribution function of the standard normal distribution. This model can be fitted using the command

```
fit.c = glm(cbind(dead,n-dead)~logdose,
          family=binomial(link=probit),data=beetle)
```

Try out this model and plot the fitted curve together with the fitted curve for the initial logit-model. Also calculate the likelihood for the two fitted models and comment.

(Hint: `logLik(fit.c)` gives the log-likelihood value.)

ADDITIONAL EXERCISE 6
Assume that for $i = 1, 2, \ldots, n$, we have

$$Y_i^* = \sum_{j=1}^{p} \beta_j^* x_{ij} + \sigma\epsilon_i,$$

where the ϵ_i's are iid with cumulative distribution (cdf) F. We further assume that the ϵ_i's have mean 0, and that their distribution is symmetric around 0, so that $F(-z) = 1 - F(z)$. We now define Bernoulli distributed random variables $Y_1, \ldots, Y_n$ by $Y_i = I\{Y_i^* > 0\}$, where $I\{\cdot\}$ is the indicator function, and let $\pi_i = P(Y_i = 1)$.

a) Show that $\pi_i = F\left(\sum_{j=1}^{p} \beta_j x_{ij}\right)$, where $\beta_j = \beta_j^*/\sigma$.

b) Which model for the π_i's do you obtain if $\epsilon_i \sim N(0, 1)$?

c) Which model for the π_i's do you obtain if $F(z) = e^z/(1 + e^z)$?

d) Which model for the π_i's do you obtain if $\epsilon_i \sim$ uniform$(-1/2, 1/2)$? You may assume that $\sum_{j=1}^{p} \beta_j x_{ij} \in (-1/2, 1/2)$.

ADDITIONAL EXERCISE 7
Consider the Poisson distribution with pmf

$$f(y; \mu) = \frac{\mu^y}{y!} \exp(-\mu), \quad y = 0, 1, 2, \ldots.$$

a) Show that the Poisson distribution is in the exponential dispersion family. That is, show that it can be written on the form

$$\exp\{[\theta y - b(\theta)]/a(\phi) + c(y, \phi)\},$$

and determine θ, $b(\theta)$, $a(\phi)$ and $c(y, \phi)$.

b) Find the mean and variance for the Poisson distribution using the relations (4.3) and (4.4) in the text book.

c) Determine the canonical link function for a Poisson GLM.

ADDITIONAL EXERCISE 8
Consider the geometric distribution with pmf

$$f(y; \pi) = \pi(1 - \pi)^y, \quad y = 0, 1, 2, \ldots.$$

a) Show that the geometric distribution is in the exponential dispersion family.

Then consider the negative binomial distribution with pmf

$$f(y;\pi) = \binom{y+r-1}{r-1}\pi^r(1-\pi)^y, \quad y = 0,1,2,\ldots,$$

where r is a given natural number.

b) Show that the negative binomial distribution is in the exponential dispersion family.

c) Find the mean and variance for the negative binomial distribution using the relations (4.3) and (4.4) in the text book.

Additional exercise 9

a) Show that the exponential distribution with pdf

$$f(y;\lambda) = \lambda\exp(-\lambda y), \quad y > 0,$$

is in the exponential dispersion family. That is, show that the pdf can be written on the form $\exp\{[\theta y - b(\theta)]/a(\phi) + c(y,\phi)\}$. Determine the relation between λ and the natural parameter θ, and use (4.3) and (4.4) in the text book to determine the mean and variance for the exponential distribution.

b) Consider the gamma distribution with pdf given as

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

Show that this is in the exponential dispersion family. Then find the mean and variance using (4.3) and (4.4) in the text book.

Additional exercise 10

The *moment generating function* is a useful tools for determining the moments of a distribution. If Y has pdf/pmf $f(y)$, the moment generating functions is given as

$$M_Y(t) = \mathrm{E}[\exp(Yt)] = \begin{cases} \sum e^{yt}f(y) & \text{if } Y \text{ is discrete} \\ \int e^{yt}f(y)dy & \text{if } Y \text{ is continuous} \end{cases}$$

The cumulant generating function is then given by $R_Y(t) = \log M_Y(t)$.

a) Show that $E(Y) = M_Y'(0)$. Also show that $M_Y(0) = 1$ and $E(Y^r) = M_Y^{(r)}(0)$. (You may assume that the order of differentiation and integration/summation may be interchanged.)

b) Show that $E(Y) = R_Y'(0)$ and $\mathrm{var}(Y) = R_Y''(0)$

Additional exercise 11

Assume that Y has a pdf/pmf from the exponential dispersion family given by (4.1) in the text book.

a) Show that the moment generating function is given by

$$M_Y(t) = \exp\{[b(\theta + ta(\phi)) - b(\theta)]/a(\phi)\}$$

and determine the cumulant generating function $R_Y(t) = \log M_Y(t)$.

b) Use the result in a) to show that $E(Y) = b'(\theta)$ and $\text{var}(Y) = b''(\theta)a(\phi)$; cf. (4.3) and (4.4) in the text book.

Additional exercise 12

At the lectures we considered data on the occurrence of "low birthweight" (i.e. less than 2.5 kg) in a sample of 189 newborn babies. The data are from the book *Applied logistic regression* by Hosmer and Lemeshow (Wiley, New York, 1989).

You may read the data into R by the commands:

```
data="http://www.uio.no/studier/emner/matnat/math/STK3100/data/lowbirthweight.txt"
lowbirthweight=read.table(data,header=T)
```

The data are organized with one line for each of the 189 babies and the following variables in the first nine columns (the tenth column is the birth weight (`bwt`), and should not be used in the analysis):

- `low`: Indicator of birth weight less than 2.5 kg (0=No, 1=Yes)
- `age`: Mother's age in years
- `lwt`: Mother's weight in pounds at last menstrual period
- `race`: Mothers race (black, other, white)
- `smoke`: Smoking status during pregnancy (0=No, 1=Yes)
- `ht`: History of hypertension (0=No, 1=Yes)
- `ui`: Presence of uterine irritability (0=No, 1=Yes)
- `ftv`: Number of physician visits during the first trimester (0=None, 1=One, 2=Two, etc.)
- `ptl`: Number of previous premature labours (0=None, 1=One, etc.)

(a) Fit a logistic regression model using `low` as response and the other eight variables given above as covariates.

(b) Perform a backwards elimination, where you remove non-significant covariates, one at a time, until all covariates in the model have a significant effect.

Hint: It is convenient to use the command `drop1`; cf. the R code from the lectures.

(*c*) Discuss what "the final model" tells you about the effect of the covariates.

Additional exercise 13 (From McCullagh and Nelder, 1989)
Assume $Y_1, ..., Y_n$ are *iid* with pdf/pmf $f_Y(y; \theta, \phi)$ from the exponential dispersion family. From additional exercise 11 we then have the the moment generating function (mgf) the Y_i's is $M_Y(t) = \exp\{[b(\theta + ta(\phi)) - b(\theta)]/a(\phi)\}$.

(*a*) Show that the arithmetic average $\overline{Y}$ also has a distribution within the exponential dispersion family.

Hint: Find the mgf of $\overline{Y}$ and remember that the mgf uniquely determines the distribution.

(*b*) Assume now $Y_1, ..., Y_n$ are *iid* from the Bin$(1, \pi)$ distribution. Find the distribution for $\overline{Y}$ in this case.

Additional exercise 14
Let $L(\boldsymbol{\beta}, \phi)$ be the log-likelihood based on n independent observations from a generalized linear model. Define (with the dispersion parameter ϕ considered fixed)

$$S_j(\boldsymbol{\beta}) = \frac{\partial}{\partial \beta_j} L(\boldsymbol{\beta}), \qquad j = 1, ..., p$$

$$\mathcal{J}_{hj}(\boldsymbol{\beta}) = E\left[-\frac{\partial^2}{\partial \beta_h \partial \beta_j} L(\boldsymbol{\beta}) \right], \qquad h, j = 1, ..., p$$

Show that

$$\mathrm{Cov}[S_h(\boldsymbol{\beta}), S_j(\boldsymbol{\beta})] = \mathcal{J}_{hj}(\boldsymbol{\beta}).$$

Use this to show that the matrix $\boldsymbol{\mathcal{J}}(\boldsymbol{\beta}) = \{\mathcal{J}_{hj}(\boldsymbol{\beta})\}$ cannot be negative definite.

Additional exercise 15
Consider the beetle data from the lectures; see also additional exercise 5. Fit a logistic regression model with log-dose as covariate. Use the fitted model to estimate LD50, i.e., the log-dose corresponding to 50% mortality. Use the result of exercise 5.10 in the text bok to compute a 95% confidence interval for LD50.

Additional exercise 16
This problem is about testing a two sided fully specified hypothesis for a scalar parameter. There are several possible tests: the Wald test, the score test, and the likelihood ratio test. In the exercise you shall compare these tests for a model where the variable is binomially distributed.

Suppose $V \sim \mathrm{bin}(n, \pi)$, i.e. binomially distributed with n trials and probability of success π.

a) Find the log likelihood $L(\pi)$, the score function $S(\pi)$, and the expected information $\mathcal{J}(\pi)$ based on V.

b) Describe the approximate distribution of the maximum likelihood estimator $\hat{\pi} = V/n$ when n is large, and show that the Wald test statistic for $H_0 : \pi = \pi_0$ is given by

$$Z_W(\pi_0) = \frac{\hat{\pi} - \pi_0}{\sqrt{\hat{\pi}(1-\hat{\pi})/n}}$$

Explain that the Wald test statistic is approximately standard normally distributed under the null hypothesis.

c) Show that the score test statistic for the same null hypothesis is given as

$$Z_S(\pi_0) = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}$$

Also explain why $Z_S(\pi_0)$ is approximately standard normally distributed under H_0.

d) Show that the likelihood ratio test statistic for $H_0 : \pi = \pi_0$ can be written

$$Z^2_{LR}(\pi_0) = 2\left[V \log\left(\frac{\hat{\pi}}{\pi_0}\right) + (n-V)\log\left(\frac{1-\hat{\pi}}{1-\pi_0}\right)\right]$$

What is the approximate distribution of $Z^2_{LR}(\pi_0)$ when n is large?

e) With $n = 100$ and $v = 30$ implement and compare the three tests for $H_0 : \pi = 0.5$.

f) With $n = 100$ and $v = 5$ implement and compare the three tests for $H_0 : \pi = 0.15$.

For each of the test statistics in questions b-d, we may obtain an approximate 95% confidence interval for π by inverting the tests. The confidence intervals are given by $\{\pi_0 : Z^2_{\cdot}(\pi_0) \leq 3.84\}$.

g) Show that when $\mathrm{SE}(\hat{\pi}) = \sqrt{\hat{\pi}(1-\hat{\pi})/n}$, the 95% confidence interval based on the Wald statistic $Z_W(\pi_0)$ is $\hat{\pi} \pm 1.96 \cdot \mathrm{SE}(\hat{\pi})$.

h) Show that the confidence interval based on the score statistic $Z_S(\pi_0)$ is the solution of a quadratic equation.
Solution: $[\hat{\pi}_L, \hat{\pi}_U] = \frac{v+1.92}{n+3.84} \pm \frac{1.96\sqrt{n\hat{\pi}(1-\hat{\pi})+3.84/4}}{n+3.84}$.

i) Compute the confidence intervals when $n = 100$ and $v = 30$. [For the interval based on the likelihood ratio statistic $Z^2_{LR}(\pi_0)$ there is no explicit formula for the end points of the interval, but you can read them off a plot of $(\pi_0, Z^2_{LR}(\pi_0))$ or find them by solving two equations numerically.]

j) Compute the confidence intervals when $n = 100$ and $v = 5$.

Additional exercise 17 (Weighted least squares)
Consider the normal linear model

$$Y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

where $\varepsilon_i \sim N(0, \sigma^2/w_i)$ with w_i known.

(*a*) Define $Y_i^* = \sqrt{w_i}\, Y_i, x_{ij}^* = \sqrt{w_i}\, x_{ij}$, and $\varepsilon_i^* = \sqrt{w_i}\, \varepsilon_i$. Show that we now can write a regression model with Y_i^* as response and x_{ij}^* as covariates, where the noise terms have constant variance.

(*b*) Show that the least squares estimate $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ is given by

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{Y}$$

where $\boldsymbol{W} = \text{diag}\{w_i\}$.

Hint: Express first $\widehat{\boldsymbol{\beta}}$ by $\boldsymbol{X}^*$ and $\boldsymbol{Y}^*$.

Additional exercise 18 (The inverse Gaussian distribution)
The inverse Gaussian distribution is given by

$$f(y) = \frac{1}{\sqrt{2\pi y^3 \sigma}} \exp\left\{ -\frac{1}{2y} \left(\frac{y-\mu}{\mu\sigma} \right)^2 \right\}, \quad y > 0$$

(*a*) Show that this distribution belongs to the exponential dispersion family and identify θ, $a(\theta)$, ϕ, and $c(y; \phi)$.

(*b*) Use general results about the exponential dispersion family to find the expectation and variance function for the distribution.

(*c*) Find the canonical link for the inverse Gaussian distribution.

(*d*) Assume now that $Y_1, ..., Y_n$ are independent variables from a GLM with the inverse Gaussian distribution as response distribution. Derive the deviance in this case.

Additional exercise 19
We will consider a data set on one-year vehicle insurance policies (de Jong & Heller: *Generalized linear models for insurance data*, Cambrigde University Press, 2008). The data set contains information on the number of claims and the claim amounts for almost 70 000 policy holders. In this exercise we will focus on the claim amounts for those who had at least one claim.

You can read the data and extract the policy holders who had at least one claim by the commands:

```
data="http://www.uio.no/studier/emner/matnat/math/STK3100/data/car.txt"
car = read.table(data,header=T,sep=",")
car0 = car[car$claimcst0>0,]
car0$agecat = as.factor(car0$agecat)
car0$gender = as.factor(car0$gender)
car0$area = as.factor(car0$area)
```

Here

- `agecat` is the driver's age category: `1 (youngest), 2, 3, 4, 5, 6`
- `gender` is `M` for malles and `F` for females
- `area` is the driver's area of residence: `A, B, C, D, E, F`

A fit using the inverse Gaussian response distribution (cf. previous exercise), a `log` link function and using driver's age, gender and area as explanatory variables can be performed through the command

```
fit = glm(claimcst0~agecat+gender+area,data=car0,
  family=inverse.gaussian(link="log"))
```

(*a*) Perform the above commands and look at the summary of the results.

(*b*) Use a Wald test to test whether gender is significant. Compare this with a likelihood ratio test.

(*c*) Assume now we want to test whether driver's age is significant. Discuss problems with performing a Wald test from the summary of `fit`. Perform instead a likelihood ratio test. What is your conclusion?

Additional exercise 20

In this exercise we will consider a special case of logistic regression where we have one binary covariate, such that $x = 1$ or $x = 0$. We denote individuals with $x = 0$ as group 0 and $x = 1$ as group 1. The data can then be written as the 2×2 table below:

	Group 1	Group 0	Total
Sick	A	B	$n_{0\bullet} = A + B$
Healthy	C	D	$n_{1\bullet} = C + D$
Total	$n_{\bullet 1} = A + C$	$n_{\bullet 0} = B + D$	$n = A + B + C + D$

Let $\pi(0)$ and $\pi(1)$ be the probability for disease in group 0 and group 1. Then

$$A \sim \text{Bin}(n_{\bullet 0}, \pi(1))$$
$$B \sim \text{Bin}(n_{\bullet 1}, \pi(0))$$

where A and B are independent of each other.

(*a*) Show that $\widehat{\pi}(1) = \frac{A}{A+C}$ and $\widehat{\pi}(0) = \frac{B}{B+D}$ are the maximum likelihood estimators of $\pi(1)$ and $\pi(0)$.

Use this to find an estimator of the odds ratio. The odds ratio is defined as

$$\text{OR} = \frac{\frac{\pi(1)}{1-\pi(1)}}{\frac{\pi(0)}{1-\pi(0)}} = \frac{\pi(1)}{\pi(0)} \frac{1-\pi(0)}{1-\pi(1)}$$

(*b*) An alternative formulation of the binomial distribution is through the canonical parameter $\theta(j) = \log(\pi(j)/(1-\pi(j))), j = 0, 1$.

Discuss why $\widehat{\theta}(j) = \log(\widehat{\pi}(j)/(1-\widehat{\pi}(j))$ is the maximum likelihood estimator of $\theta(j)$.

By using standard likelihood theory, show that an estimator for the variance of $\widehat{\theta}(0)$ is given by

$$\widehat{\text{var}}[\widehat{\theta}(0)] = \frac{1}{B} + \frac{1}{D}$$

and find a similar expression for $\widehat{\text{var}}[\widehat{\theta}(1)]$.

(*c*) Show that

$$\widehat{\text{var}}[\log \widehat{OR}] = \text{SE}^2 = \frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}$$

Use this to show that

$$\widehat{\text{OR}} \cdot \exp(\pm 1.96\text{SE})$$

is an approxemately 95% confidence interval for OR.

The data underneath is from a health survey in Nord-Trøndelag and shows the number of people with and without diabetes II according to gender. The survey is from 1985 and 1995 where 38676 people where examined.

Table 1. Number of people with and without diabetes by gender.

Gender	Male	Female	Total
Diseased	377	336	713
Healthy	17864	20099	37963
Total	18241	20435	38676

(*d*) Use a chi-squared test for 2×2 tables to see if the occurrence of diabetes is different for men and women.

(*e*) Estimate the odds ratio (OR) for diabetes between men and women and calculate a 95% confidence interval for OR.

(*f*) Explain why OR = 1 is equivalent to say that the probability for disease is the same for the two groups. Use this to test if the occurrence of diabetes is different in the two groups.

(*g*) Do similar calculations in `R` using the glm-procedure and compare the results.

Hint: Make a dataframe with two rows where the first row is data for men, the second for women and an aditional column indicating gender.

Additional exercise 21 (Exam STK3100 fall 2007, problem 2)
In this problem we will consider the risk of death in the so-called post neonatal period, from the 28th day of life until the first birthday. We will only consider the sudden infant death syndrome, SIDS, and we will model the probability for a SIDS death in the period, given that the child does not die from another cause, using logistic regression.

In particular, we will consider how SIDS-deaths are related to year of birth, gender and weight at birth. 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 at birth (`vekt`) is used as a continuous covariate measured in kilograms.

a) Below is given a deviance table from a R-output with some quantities marked with "?". Fill out these 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 given R-output when only the main effects of the covariates birth weight, gender and year of birth is included. For year of birth a corner point parametrization is used, where 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
```

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

$$\begin{aligned}\pi_{ij} &= \frac{\exp(\mathbf{x}_i\boldsymbol{\beta}_j)}{1+\sum_{k=1}^{J}\exp(\mathbf{x}_i\boldsymbol{\beta}_k)},\ j=1,\ldots,J\\ \pi_{i0} &= \frac{1}{1+\sum_{k=1}^{J}\exp(\mathbf{x}_i\boldsymbol{\beta}_k)},\end{aligned}$$

where $\boldsymbol{\beta}_j$ is a vector of regression parameters. Let SIDS correspond to $j=1$.

c) 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 this logistic regression model look like.

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

Additional exercise 22

We will consider the example from the lectures with the rate of new of lung cancer cases among males in four Danish cities. You may read the data into `R` by the commands:

```
city=rep(1:4,each=5)
age=rep(1:5,times=4)
cases=c(11,11,11,10,11,13,6,15,10,12,4,8,7,11,9,5,7,10,14,8)
number=c(3059,800,710,581,509,2879,1083,923,834,634,3142,
         1050,895,702,535,2520,878,839,631,539)
```

The coding of the variables are as follows:

- `city`: City (1=Fredericia; 2=Horsens; 3=Kolding; 4=Vejle).
- `age`: Age group (1=40-54 years; 2=55-59 years; 3=60-64 years; 4=65-69 years; 5=70-74 years).
- `cases`: Number of new male lung cancer cases in the period 1968-1972.
- `number`: Approximate number of male inhabitants.

At the lectures we considered a model with main effects of age group and city. You may fit this model by the commands:

```
fit1=glm(cases~offset(log(number))+factor(age)+factor(city),family=poisson)
summary(fit1)
```

a) Why do we include an offset when we fit the model?

b) Interpret the estimates of the model.

Fredericia is an industrial city, while the other cities have less industry. It may therefore be reasonable to group the cities in two groups, with Fredericia in one group and the three other cities in another group (so we assume that the lung cancer incidence is the same in Horsens, Kolding, and Vejle).

c) Fit such a model, and compare the model fit with the model in question b.

d Estimate the rate ratio of new lung cancer cases in Fredericia compared to the three other cities. Also give a 95% confidence interval for the rate ratio.

We have so far considered age group as a categorical covariate. Another possibility is to treat age as a numeric covariate, where we for each age group use the midpoint of the group.

e) Fit a model with age as a numeric covariate, and compare the model fit with the model in question c. Interpret the estimate of age in this model.

Additional exercise 23
Let X and Y be two random variables.

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

b) Show that $\text{var}(Y) = \text{var}(E[Y|X]) + E[\text{var}(Y|X)]$.

Additional exercise 24 (Exam STK3100, 2009, problem 2, modified)
In a German survey, 1100 women were asked about the number of visits to a physician during the last three month (`numvisit`) and about various covariates that may affect the number of visits. Among these covariates we shall only consider an indicator for self reported bad health (`badh`) and age in years (`age`).

a) Below are given the results for a fit of a Poisson regression model with log link, where the response is `numvisit` and the covariates are `badh` and `age`.

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

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

```
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
---
(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 to find estimates of the rate ratios and 95% confidence intervals for visit to a physician for

(i) women with self reported bad health and self reported good health,

(ii) women of age 50 years and women of age 40 years.

Also estimate the rate for visits to a physician 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 formulation allows for overdispersion compared to the Poisson model in part a) via the specification $\text{var}(Y) = \phi\mu$, where Y is the response (the number of visits to physicians) and μ is the expectation of Y. Explain how you may estimate the effects of the covariates using a quasi-likelihood approach. Also explain how you may estimate the overdispersion parameter ϕ.

Additional exercise 25

The data set `Orthodont` is available in R from the package `nlme`. Use the command `library(nlme)` to attach the library, and the command `help(Orthodont)` to get a description of the data set.

The data are grouped after the variable `Subject`. The response variable is `distance` with the two covariates `age` and `Sex`.

In this exercise we will look at how linear mixed models can be used to analyse this data set. Before we do any analyses, we transform `Subject` to the values 1-27 by the command

```
Orthodont$ID = as.factor(as.numeric(Orthodont$Subject))
```

(a) Do some explorative analyses to become familiar with the data set.

(b) Plot `distance` against `age` for each of the subjects (in the same plot). Comment on the plot.

(c) Give the commands

```
fit1 = lme(distance ~ age,data=Orthodont,random=~1|ID)
plot(Orthodont$age,Orthodont$distance,col=Orthodont$ID)
abline(fit1$coef$fixed,lwd=4)
for (i in 1:27)
{
abline(cbind(fit1$coef$fixed[1]+fit1$coef$random$ID[i,],
       fit1$coef$fixed[2]),col=i)
}
```

Comment on the results.

(d) What are the estimates of σ_u^2 and σ_ϵ^2?

(e) Now add `sex` to the model and see if it is a significant covariate based on the Wald test. Interpret the estimates of the model.

Additional exercise 26

In this exercise we will take a closer look on REML estimation in connection with linear normal models. We assume that $\boldsymbol{Y}$ is a N-dimensional random vector and that $\boldsymbol{Y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{V})$, where the model matrix $\boldsymbol{X}$ is a $N \times p$ matrix of full rank and the covariance matrix $\boldsymbol{V}$ depends on a parameter vector $\boldsymbol{\theta}$, i.e. $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\theta})$. Note that this covers the situation with clustered data given by (9.9) in the book by Agresti [where $N = nd$ and $\boldsymbol{V}$ is given by the formula just below (9.9)].

For REML-estimation of $\boldsymbol{V}(\boldsymbol{\theta})$ we consider the linear transformation $\boldsymbol{LY}$, where

$$\boldsymbol{L} = \boldsymbol{I}_N - \boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{I}_N - \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}},$$

where $\boldsymbol{I}_N$ is the $N \times N$ identity matrix; c.f. sections 2.2.2 and 9.3.3 in the book by Agresti.

Now $\boldsymbol{LY}$ has a singular multivariate normal distribution. To obtain a non-singular distribution we omit the last p elements of the vector $\boldsymbol{LY}$, i.e. we consider $\boldsymbol{Y}^* = \boldsymbol{AY}$, where

$$\boldsymbol{A} = \left(\boldsymbol{I}_{N-p}\,,\, \boldsymbol{0}_{(N-p)\times p} \right) \boldsymbol{L}$$

is a $(N-p) \times p$ matrix of full rank.

a) Show that $\boldsymbol{AX} = \boldsymbol{0}$.

b) Show that the distribution of $\boldsymbol{Y}^* = \boldsymbol{AY}$ does not depend on $\boldsymbol{\beta}$ and specify the distribution.

c) Show that the log-likelihood based on $\boldsymbol{Y}^*$ may be written

$$L(\boldsymbol{\theta}) = -\tfrac{N-p}{2}\log(2\pi) - \tfrac{1}{2}\log|\boldsymbol{A}\boldsymbol{V}(\boldsymbol{\theta})\boldsymbol{A}^{\mathrm{T}}| - \tfrac{1}{2}\boldsymbol{Y}^{*\mathrm{T}}\left[\boldsymbol{A}\boldsymbol{V}(\boldsymbol{\theta})\boldsymbol{A}^{\mathrm{T}}\right]^{-1}\boldsymbol{Y}^*$$

Let $\widehat{\boldsymbol{\theta}}$ be the value of $\boldsymbol{\theta}$ that maximizes the log-likelihood in c Then the REML estimator of the covariance matrix is given as $\boldsymbol{V}(\widehat{\boldsymbol{\theta}})$.

d) Assume that $\widetilde{\boldsymbol{A}} = \boldsymbol{B}\boldsymbol{A}$, where $\boldsymbol{B}$ is a nonsingular matrix of dimension $(N-p) \times (N-p)$. Show that the REML estimator based on $\widetilde{\boldsymbol{A}}$ is identical with the REML estimator based on $\boldsymbol{A}$.

(*Hint:* Show that the log-likelihood based on $\boldsymbol{Y}^*$ and the log-likelihood based on $\widetilde{\boldsymbol{Y}} = \widetilde{\boldsymbol{A}}\boldsymbol{Y}$ only differ by a constant.)

We now assume that the covariance matrix takes the form $\boldsymbol{V}(\boldsymbol{\theta}) = \sigma^2 \boldsymbol{V}_0$, where $\boldsymbol{V}_0$ is a known matrix.

e) Show that the REML estimator for σ^2 is given by

$$\widehat{\sigma}^2_{\text{REML}} = \frac{1}{n-p} \boldsymbol{Y}^{*\text{T}} \left[\boldsymbol{A}\boldsymbol{V}_0\boldsymbol{A}^{\text{T}} \right]^{-1} \boldsymbol{Y}^*$$

f) Show that the estimator in question e is unbiased.

(*Hint:* Use (2.7) in the book by Agresti.)

Additional exercise 27 (Exam STK3100, 2008, problem 1, modified)

a) Show that the probability mass function of a Poisson distributed variable can be written as $f(y;\theta) = \exp\{y\theta - b(\theta) + c(y)\}$. What is the relation between the expectation μ of the Poisson distributed variable and the parameter θ. Find explicit expressions for $b(\theta)$ and $c(y)$.

b) Suppose that $Y_1, \ldots, Y_n$ are independent Poisson distributed variables with expectations $\mu_i = \exp(\beta_0 + \beta_1 x_i)$, for $i = 1, \ldots, n$, where β_0 and β_1 are regression parameters and x_i are known covariates. Show that this is a particular case of generalized linear models (GLM).

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

$$\boldsymbol{U}(\beta_0, \beta_1) = \begin{pmatrix} U_0(\beta_0, \beta_1) \\ U_1(\beta_0, \beta_1) \end{pmatrix} = \sum_{i=1}^{n} \begin{pmatrix} 1 \\ x_i \end{pmatrix} (Y_i - \mu_i).$$

Also find an expression for the expected information matrix.

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

e) Use the result in question d to find an expression for the deviance in a generalized linear model with responses that 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 events Y_i for each individual, but the analysis is based on variables indicating the occurrence of at least one event, i.e. $Y_i' = I(Y_i > 0)$. Let $\pi_i =$

$P(Y_i' = 1)$. Show that this is a generalized linear model for binary data with the same linear predictor as in question b and with link function

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

What is this link function called?

g) Below are given the results from an analysis of vehicle insurance policies, where the response is whether an insured has reported one or more accidents last year. The model for the observations is a Poisson regression model as described in question b, but in the analysis the binary responses corresponding to the model developed in question 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 numeric covariate 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 below to compute the rate ratios for 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) Also find 95% confidence intervals for the rate ratio 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 ,0000 dollars

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

Additional exercise 28

We assume that $Y_1, \ldots, Y_n$ are independent and normally distributed with common variance σ^2. The Y_i's are observations from three groups, so we have

$$\begin{aligned} E(Y_1) = E(Y_2) = \cdots = E(Y_{n_1}) &= \beta_1, \\ E(Y_{n_1+1}) = E(Y_{n_1+2}) = \cdots = E(Y_{n_1+n_2}) &= \beta_2, \\ E(Y_{n_1+n_2+1}) = E(Y_{n_1+n_2+2}) = \cdots = E(Y_n) &= \beta_3, \end{aligned} \tag{3}$$

where $n = n_1 + n_2 + n_3$. We introduce the vectors $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^{\mathrm{T}}$ and $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)^{\mathrm{T}}$, and let $\mathbf{0}_m$ and $\mathbf{1}_m$ be m-vectors of 0's and 1's.

a) Show that model (3) may be given on vector/matrix form as

$$\boldsymbol{\mu} = E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta},$$

where

$$\boldsymbol{X} = \begin{bmatrix} \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \mathbf{0}_{n_1} \\ \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \mathbf{0}_{n_2} \\ \mathbf{0}_{n_3} & \mathbf{0}_{n_3} & \mathbf{1}_{n_3} \end{bmatrix},$$

and $\boldsymbol{\beta} = (\beta_1, \beta_1, \beta_3)^{\mathrm{T}}$. What is the rank of the model matrix?

b) Show that the projection matrix onto the model space $C(\mathbf{X})$ may be given as

$$\boldsymbol{P}_1 = \begin{bmatrix} \frac{1}{n_1}\mathbf{1}_{n_1}\mathbf{1}_{n_1}^{\mathrm{T}} & \mathbf{0}_{n_1} & \mathbf{0}_{n_1} \\ \mathbf{0}_{n_2} & \frac{1}{n_2}\mathbf{1}_{n_2}\mathbf{1}_{n_2}^{\mathrm{T}} & \mathbf{0}_{n_2} \\ \mathbf{0}_{n_3} & \mathbf{0}_{n_3} & \frac{1}{n_3}\mathbf{1}_{n_3}\mathbf{1}_{n_3}^{\mathrm{T}} \end{bmatrix}.$$

c) Determine the fitted values $\widehat{\boldsymbol{\mu}} = \boldsymbol{P}_1\boldsymbol{Y}$.

If $\beta_1 = \beta_2 = \beta_3$ we have the null model where all the Y_i's have the same means. The projection matrix for the null model is known to be $\boldsymbol{P}_0 = n^{-1}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}$. We then have the orthogonal decomposition

$$\boldsymbol{Y} = \boldsymbol{P}_0\boldsymbol{Y} + (\boldsymbol{P}_1 - \boldsymbol{P}_0)\boldsymbol{Y} + (\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{Y}$$

with corresponding sum of squares decomposition

$$\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y} = \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{P}_0\boldsymbol{Y} + \boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{P}_1 - \boldsymbol{P}_0)\boldsymbol{Y} + \boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{Y}.$$

c) Let $\overline{Y}_1 = n_1^{-1}\sum_{i=1}^{n_1} Y_i$, $\overline{Y}_2 = n_2^{-1}\sum_{i=n_1+1}^{n_1+n_2} Y_i$, and $\overline{Y}_3 = n_3^{-1}\sum_{i=n_1+n_2+1}^{n} Y_i$ denote the means in the three groups, and let $\overline{Y} = n^{-1}\sum_{i=1}^{n} Y_i$ be the overall mean. Show that

$$\boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{P}_1 - \boldsymbol{P}_0)\boldsymbol{Y} = \sum_{j=1}^{3} n_j(\overline{Y}_j - \overline{Y})^2$$

and that

$$\boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{Y} = \sum_{i=1}^{n_1}(Y_i-\overline{Y})^2 + \sum_{i=n_1+1}^{n_1+n_2}(Y_i-\overline{Y})^2 + \sum_{i=n_1+n_2+1}^{n}(Y_i-\overline{Y})^2 \stackrel{\text{def}}{=} \text{SSE}$$

d) Explain that

$$\sum_{j=1}^{3} n_j(\overline{Y}_j - \overline{Y})^2/\sigma^2 \qquad \text{and} \qquad \text{SSE}/\sigma^2$$

are independent and determine their distributions.

e) Derive an F-statistic for testing the null hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3$ versus the alternative hypothesis that not all the β_j's are equal. Determine the distribution of the test statistic under H_0 and under H_A.