Exercises in STK3100/4100.

The datasets needed in the exercises are available from the package STK3100. This package can be downloaded from the homepage of the course and installed using the command `install.packages("STK3100_1.0.tar.gz", repos=NULL, type="source")` in R. The command `library(STK3100)` will load the package and `data()` will list the available datasets. For further info about the datasets use the command `help("name of dataset")`.

Exercise 1
Read chapter 4 in Heller & Jong. Further into the course this material will be assumed known.

Exercise 2 (Linear models and testing of several coefficients simultaneously)
We study the dataset birthweight from the package `STK3100`. Use `data(birthweight)` to load the data set in R. We assume the model is

$$Y_{jk} = \alpha_j + \beta x_{jk} + \varepsilon_{jk}$$

where j is gender and k is the index of baby in each group of gender. The Y's are saved in a single vector with associated covariates. For the analysis you may use commands below

```
birthweight$sex = as.factor(birthweight$sex)
lm(vekt~sex+svlengde-1,data=birthweight)
```

An alternative model is when the slope also depends on sex:

$$Y_{jk} = \alpha_j + \beta_j x_{jk} + \varepsilon_{jk}$$

This model can be formulated in various ways. Here we will define

```
birthweight$x4 = birthweight$svlengde*(birthweight$sex==1)
birthweight$x5 = birthweight$svlengde*(birthweight$sex==2)
```

and then fit the model by

```
lm(vekt~sex+x4+x5-1,data=birthweight)
```

(*a*) Try out this model and ensure that you understand how the model is formulated.

(*b*) We will know test to see if length of pregnancy is relevant for the modeling of birthweights. This corresponds to test whether $H_0 : \beta_1 = \beta_2 = 0$. Use the theory of section 4.15 in de Jong & Heller to perform the F-test. This is a special case of the more general problem in section 4.15.

EXERCISE 3

In class we discussed the model

$$\log(\pi_i/(1-\pi_i)) = \beta_0 + \beta_1 x_i$$

for the beetle data (where x_i = poisoning dose and π_i is the probability of killing beetles in group i).

(a) Try out the alternative model

$$\log(\pi_i/(1-\pi_i)) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$$

on the beetle data located in the `STK3100`-package. (Note that you in R have to use `I(x^2)` on the quadratic term)

Plot the fitted curve together with the curve from the initial linear model.

Consider if x_i^2 is significant by checking the related P-value.

(b) By using that $\hat{\boldsymbol{\beta}} \approx N(\boldsymbol{\beta}, \boldsymbol{I}^{-1})$, find correlations between the different β-estimates. (Here $\boldsymbol{I}$ is the information matrix, we will discuss the validity of this statement later on in the course, as well as defining $\boldsymbol{I}$ more properly.)

Why are the correlations that strong?

Given this correlation structure, consider the procedure of testing the relevance of the quadratic term in the previous subtask.

(Hint: Use the covariance matrix of $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)'$. This can be found by the command `summary(glmfit)$cov.scaled` when `glmfit` is the fitted glm-model with quadratic term.)

(c) An alternative link-function for the logistic is the *probit*-function given by

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

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

```
glmfit2 = glm(cbind(dead,tot-dead)~dose,family=binomial(link=probit))
```

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(glmfit2)` gives the log-likelihood value.)

Exercise 4

Moment generating functions are useful tools for determining the means. For Y with density function $f(y)$, we define moment generating functions as

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

(a) Show that

$$E[Y] = M_Y'(0)$$

Also show that

$$E[Y^r] = M_Y^{(r)}(0)$$

and $M_Y(0) = 1$.

(Hint: In the continuous case, we must assume that it is possible to switch derivation and integration)

(b) Estimate the moment generating function for the Poisson distribution and use this to show that the mean and variance are equal.

(c) Estimate the moment generating function for the Exponential distribution and use this to find the mean and variance.

(d) Estimate the moment generating function for the Gamma distribution (use the notation in de Jong & Heller 2.7) and use this to show the formulas for mean and variance.

Exercise 5

In class we saw that the probability mass function for the Poisson distribution, the binomial distribution and the probability density function for the normal distribution with variance $\sigma^2 = 1$ can be written on the form $f(y;\theta) = c(y)\exp(\theta y - a(\theta))$.

a) Find (preferably without looking at the lecture notes) θ as a function of the initial parameter in these distributions. Also find the functions $a(\theta)$ and $c(y)$.

b) We also saw that the moment generating function $M_Y(t) = \exp(a(\theta + t) - a(\theta))$. Use this to find the moment generating function for the three distributions.

The normal distribution with variance σ^2 is also in the exponential family with dispersion parameter ϕ. Then the pdf can be written on the form $f(y;\theta) = c(y,\phi)\exp((\theta y - a(\theta))/\phi)$. The moment generating function for the exponential class with dispersion parameter is $M_Y(t) = \exp((a(\theta + t\phi) - a(\theta))/\phi)$.

c) Find ϕ and the moment generating function for $Y \sim \mathrm{N}(\mu, \sigma^2)$ with the initial parametrization.

Exercise 6

a) Show that the exponential distribution with pdf $\lambda \exp(-\lambda y)$ is in the exponential family. That is, show that it can be written on the form $c(y)\exp(\theta y - a(\theta))$. Find mean and variance for the exponential distribution.

b) Show that that the pdf for the gamma distribution

$$f(y) = \frac{y^{\nu-1}\lambda^{\nu}}{\Gamma(\nu)}\exp(-\lambda y)$$

for a given parameter ν also can be written on this form. Then find mean and variance for a gamma distributed variable based on what you have found.

c) Show that the gamma distribution can be written as an exponential class with dispersion parameter.

Hint: write the density as

$$f(y) = \frac{1}{\Gamma(\nu)}\left(\frac{\nu}{\mu}\right)^{\nu} y^{\nu-1}\exp(-\frac{\nu}{\mu}y)$$

Exercise 7

a) Show that the geometric distribution with pmf $\pi(1-\pi)^y$ for $y = 0, 1, \ldots$ can be written on the form $c(y)\exp(\theta y - a(\theta))$. Find the mean and variance for geometric distributed variables.

b) Show that the negative binomial distribution with pmf $f(y) = \binom{y+r-1}{r-1}\pi^r(1-\pi)^y$ for $y = 0, 1, \ldots$ also can be written in this way.

c) Suppose $Y|\lambda \sim \text{Po}(\lambda)$ (Poisson-distributed) where λ is a gamma distributed variable with pdf as in exercise 6b. Find the marginal distribution for Y and discuss why this also is the negative binomial distribution.

d) Find mean and variance for negative binomial variables.

Exercise 8 (From McCullagh and Nelder, 1989)
Assume $f_0(y)$ is a probability density function (or probability mass function in the discrete case) with moment generating function

$$M(t) = E[\exp(tY)] = \exp(a(t))$$

assumed to be finite for an interval including 0. Consider now the exponential weighted density

$$f_Y(y;\theta) \propto \exp(\theta y) f_0(y).$$

(*a*) Find the normalizing constant for $f_Y(y;\theta)$ and show that you get a distribution within the exponential family.

(*b*) Assume now that y is a discrete variable and $f_0(y) \propto 1/y!$. Find $f_Y(y;\theta)$ and show that this is a well-known distribution.

Exercise 9 (From McCullagh and Nelder, 1989)
Assume $Y_1, ..., Y_n$ is iid with density $f_Y(y;\theta,\phi)$ within the exponential family.

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

Hint: Show first that $\sum_i Y_i$ has a distribution within the exponential family by the use of the moment generating function. Thereafter transform to the average.

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

Exercise 10
Let $l(\boldsymbol{\beta},\phi)$ be the log-likelihood based on n independent observations within the generalized linear model. Define (with ϕ considered fixed)

$$s_j(\boldsymbol{\beta},\phi) = \frac{\partial}{\partial\beta_j} l(\boldsymbol{\beta},\phi), \qquad j = 0, ..., p$$

$$I_{j,k}(\boldsymbol{\beta},\phi) = E\left[-\frac{\partial^2}{\partial\beta_j\beta_k} l(\boldsymbol{\beta},\phi)\right], \qquad j,k = 0, ..., p$$

Show that

$$\text{Cov}[s_j(\beta,\phi), s_k(\beta,\phi)] = I_{j,k}(\beta,\phi).$$

Use this to show that the matrix $\boldsymbol{I}(\boldsymbol{\beta},\phi) = \{I_{j,k}(\beta,\phi)\}$ cannot be negative definite.

Exercise 11
We define a p-dimensional vector

$$\mathbf{Y} = (Y_1, \ldots, Y_p)'$$

to be multivariate Gaussian distributed if we can write

$$\mathbf{Y} = \mathbf{AZ} + \boldsymbol{\mu}$$

where $\mathbf{Z}' = (Z_1, \ldots, Z_p)$ is a vector of p independent $N(0,1)$ variables Z_i, $\boldsymbol{\mu}' = (\mu_1, \ldots, \mu_p)$ is an arbitrary p-dimensional vector of numbers and $\boldsymbol{A}$ is a non-singular matrix of dimension $p \times p$.

(*a*) Find the expectation vector and the covariance matrix of $\mathbf{Y}$.

It can be shown that the multivariate Gaussian distribution is uniquely defined by its expectation vector and covariance matrix, but this you do not need to show.

(*b*) Let $\boldsymbol{B}$ be a non-singular $p \times p$ matrix. Show that $\mathbf{V} = \mathbf{BY}$ also is multivariate Gaussian distributed and find its expectation vector and covariance matrix.

Define $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^T$. Then $\boldsymbol{\Sigma}$ is positive definite and there exists an lower-triangular matrix $\boldsymbol{L}$ such that $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^T$.

(*c*) Show that $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{Z}+\boldsymbol{\mu}$ and $\boldsymbol{Y}_2 = \boldsymbol{L}\boldsymbol{Z}+\boldsymbol{\mu}$ have identical distributions.

(*d*) Let $\boldsymbol{B}$ be a $q \times p$ matrix with $q \leq p$ and $B_{i,j} = 1$ if $j = i$ and $= 0$ otherwise.

Show that $\boldsymbol{V} = \boldsymbol{B}\boldsymbol{Y}$ is multivariate Gaussian distributed.

(*e*) Let now $\boldsymbol{B}$ be a $q \times p$ matrix av rank $q \leq p$. Show that $\mathbf{V} = \mathbf{BY}$ is multivariate Gaussian distributed also in this case.

Hint: Extend $\boldsymbol{B}$ with $p-q$ rows that are orthogonal to the first q rows.

Exercise 12 (**R** exercise)
This exercise considers the beetle data that we have discussed in the lectures. Parts of the exercise can be solved by copying R commands given there.

(*a*) Try out logistic regression on the data. Calculate the log-likelihood value for the fitted model by using the command `logLik` in R.

(*b*) Try out probit regression. Calculate the log-likelihood value also in this case.

(*c*) Do similarly using the complementary log-log link.

(*d*) Plot the data and the fitted values for the three models. Comment on the results in relation to the log-likelihood values you obtained.

Exercise 13 (**R** exercise)
Repeat the previous exercise but now with Poisson regression on the data giving number of previous children related to age of pregnant mother. Try out different link-functions.

Hint: Look at `help(poisson)` in **R** to see the options available for link-functions.

Exercise 14 (Weighted least squares and GLM-fitting)
We will in this exercise look at a connection between weighted least squares and the Fisher-scoring algorithm for GLM's.

Consider first a linear model

$$y_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

where $x_{i0} = 1$ and $\varepsilon_i = N(0, \sigma^2/w_i)$. Here w_i signals the precision in observation i.

(a) Define $y_i^* = \sqrt{w_i} y_i, x_{ij}^* = \sqrt{w_i} x_{ij}, \varepsilon_i^* = \sqrt{w_i}\varepsilon_i$. 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 $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{y}$$

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

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

We will now turn to GLM's, and we remember that the elements in the score function $\boldsymbol{s}(\boldsymbol{\beta})$ and the expected information matrix $\boldsymbol{I}(\boldsymbol{\beta})$ are given by

$$s_j(\boldsymbol{\beta}) = \frac{1}{\phi} \sum_{i=1}^{n} x_{ij} \frac{y_i - \mu_i}{g'(\mu_i) V(\mu_i)}, \quad I_{j,k}(\boldsymbol{\beta}) = \frac{1}{\phi} \sum_{i=1}^{n} \frac{x_{ij} x_{ik}}{g'(\mu_i)^2 V(\mu_i)}$$

where the μ_i's are indirectly specified through $\boldsymbol{\beta}$.

(c) Show that by defining $\boldsymbol{X} = \{x_{ij}\}$, $\boldsymbol{y} = (y_1, ..., y_n)^T$, $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\beta}) = (\mu_1, ..., \mu_n)^T, \boldsymbol{G}(\boldsymbol{\beta}) = \text{diag}\{g'(\mu_i)\}$ and $\boldsymbol{W}(\boldsymbol{\beta}) = \text{diag}\{1/g'(\mu_i)^2 V(\mu_i)\}$ that

$$\boldsymbol{s}(\boldsymbol{\beta}) = \frac{1}{\phi} \boldsymbol{X}^T \boldsymbol{W}(\boldsymbol{\beta}) \boldsymbol{G}(\boldsymbol{\beta}) (\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\beta})), \quad \boldsymbol{J}(\boldsymbol{\beta}) = \frac{1}{\phi} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}.$$

(d) Show that the Fisher scoring algorithm can be written as

$$\boldsymbol{\beta}^{(k+1)} = (\boldsymbol{X}^T \boldsymbol{W}(\boldsymbol{\beta}^{(k)}) \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W}(\boldsymbol{\beta}^{(k)}) \boldsymbol{z}^{(k)}$$

where

$$\boldsymbol{z}^{(k)} = \boldsymbol{X}\boldsymbol{\beta}^{(k)} + \boldsymbol{G}(\boldsymbol{\mu}(\boldsymbol{\beta}^{(k)}))(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\beta}(^{(k)}))$$

and use this to explain how weighted least squares can be used to update $\hat{\boldsymbol{\beta}}$.

(*e*) Assume now that $Y_i \sim N(\mu_i, \sigma^2)$ with $\mu_i = \sum_{j=0}^{p} \beta_j x_{ij}$. Show that the Fisher scoring algorithm converges in the 1. iteration to the least squares estimator $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$.

Exercise 15 (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 family and identify θ, ϕ, $a(\theta)$ and $c(y;\phi)$.

(*b*) Use general results about the exponential 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.

Consider now the vehicle claim data set from de Jong & Heller. We will be interested in modeling claim size (restricted to those claims having a positive claim). You can read the data (available from the course home page) by

```
car = read.table("car.txt",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)
```

where the second command pick out the positive claims. A fit using the inverse Gaussian response distribution, 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"))
```

(*e*) Perform these commands and look at the summary of the results.

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

(g) Assume now we want to test whether driver's age is significant. Discuss problems with performing a direct Wald test from the summary of `fit` directly.

Perform instead a likelihood ratio test. What is your conclusion?

EXERCISE 16
Show that the deviances of the normal, Poisson and binomial distributions are as given below:

- Normal distribution: $\Delta = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \mu_i)^2$
- Poisson distribution: $\Delta = 2\sum_{i=1}^{n}[Y_i \log(Y_i/\mu_i) - (Y_i - \mu_i)]$
- Binomial distribution:
  $\Delta = 2\sum_{i=1}^{n}[Y_i \log(Y_i/(n_i\pi_i)) + (n_i - Y_i)\log((n_i - Y_i)/(n_i(1-\pi_i)))]$

EXERCISE 17
In a saturated (full) model we have one parameter for each observation. The saturated model gives the highest possible log-likelihood ($\tilde{l}$) compared to all possible models. The deviance is then defined as $\Delta = 2[\tilde{l} - l]$ where l is the log-likelihood for any given model.

We have that Y_i is a independent negative binomial distribution with pmf

$$f(y) = \binom{y+r-1}{r-1}\pi_i^r(1-\pi_i)^y \text{ for } y = 0, 1, \ldots.$$

a) Find the estimates $\tilde{\pi}_i$ for π_i in the saturated model.

b) Express $\tilde{l}$ with the $\tilde{\pi}_i$'s.

c) Express the deviance Δ by $\tilde{\pi}_i$'s and π_i's.

d) Look at the situation where $y_i = 0$. Check to see if the expressions in b) and c) still holds if we define $0\log 0 = 0$.

EXERCISE 18
In this exercise we will look at some data from Baxter et al. (1980) *Transactions 21 Congress of Actuaries* **2-3**, 11-29. The dataset includes claims in a portfolio of insured cars from an English insurance company. Number of claims are registered divided in three classification factors with four levels each:

- Age of policyholder:
  1 = below 25
  2 = 25-29
  3 = 30-35
  4 = over 35

- Engine volume in liter:
  1 = below 1
  2 = 1-1,5
  3 = 1,5-2
  4 = over 2
- District:
  4 = London and other large cities.
  1–3 = other districts.

The dataset is available from the package `STK3100`. Use `data(claims)` to load the dataset and `help(claims)` for further info. You will also need to change the classes for the different columns. The first three should be factors and the last two numeric. You can use the commands below

```
library(STK3100); data(claims)
claims$alder <- as.factor(claims$alder)
claims$motorvolum <- as.factor(claims$motorvolum)
claims$distrikt <- as.factor(claims$distrikt)
claims$antforsikret <- as.numeric(claims$antforsikret)
claims$antskader <- as.numeric(claims$antskader)
```

a) What assumptions on the distribution are reasonable with respect to number of claims? How can you model the significance of policyholders age, engine volume and district?

b) Perform an analysis on the dataset that clarifies the significance of age, engine volume, district and any potential interactions between these.

c) At this point the covariates age, engine volume and district are defined as factors. See if there is a linear trend in age and engine volume by fitting a model where these factors are added as numeric covariates. Is it possible to make a simplification in the model from the effect of district? (i.e. can it me fitted with fewer parameters?)

d) Perform an analysis on the residuals in the final model. Is there something about the residuals that suggests that the model is not satisfactory?

e) Interpreter the estimates from the model in c) as rate-ratios.

   Hint: look at the slides, lecture 5

f) Estimate the rate of an insured person in age category 25-29 years old, car with engine volume 1,5-2 liter and living in London. Also estimate a 95% confidence interval for this rate.

Exercise 19

In this exercise we look at a special case of logistic regression where we have one binary covariate, such that $x \in \{0, 1\}$. We denote individuals with $x = 0$ as group 0 and $x = 1$ as group 1. We also let Y be an indicator of disease. The data can then be written as the 2x2 table below:

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

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

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

where A and B are independent of each other.

(a) Show that $\hat{\pi}(1) = \frac{A}{A+C}$ and $\hat{\pi}(0) = \frac{B}{B+D}$ are ML estimates for $\pi(1)$ and $\pi(0)$.

Use this to find an estimate for 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 $\hat{\theta}(j) = \log(\hat{\pi}(j)/(1 - \hat{\pi}(j))$ is an ML estimate for $\theta(j)$.

By using standard likelihood theory, show that

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

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

(c) Show that

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

Use this to show that

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

is an approximately 95% CI for OR.

The data underneath is from a health survey in Nord-Trøndelag and shows the number of people with and without diabetes II divided by gender. The survey is from 1985 and 1995 where 38676 people where examined. In the table we see how diabetes is related with gender.

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 two-sample binomial test to see if the occurrence of diabetes is different for men and women (see chi-squared test: 13.1 in Devore & Berk).

(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 from men, the second from women and an additional column explaining gender.

(h) Use the new table underneath based om BMI and repeat the calculations of OR for diabetes with BMI$<$ 25 and BMI$\geq$ 25 (BMI (Body-Mass Index) Calculated: w/h^2 where w is weight in kilo and h er height in meter).

Table 2. Number of people with and without diabetes against BMI over and under 25.

BMI	< 25	≥ 25	Total
Diseased	90	623	713
Healthy	21689	16274	37963
Total	21779	16897	38676

Exercise 20

The dataset `sau` contains weight of lambs at slaughter from a Norwegian county collected in the years 1989-1998. We have the following 7 variables:

- `hvkt` Body weight for lambs at slaughter (Response variable)
- `aar` year
- `agewe` Age of mother
- `kjoenn` Gender
- `burdH` Number of lambs in the litter
- `alderlam` Age of lamb at slaughter (in days)
- `NAO` a climate index for the current year.

The objective is to explain the variation in the data based on the available covariates. The dataset is available from the package `STK3100`.

(a) Do some exploratory analysis to become familiar with the data set.

(b) Try out a generalized linear model with all covariates, gamma distribution and log-link.

What is the (residual) deviance for the model? (remember that R does not count for the dispersion parameter)

(c) Do a similar analysis, but with the inverse Gaussian distribution and log-link

Compare the two models using AIC-criterion.

(d) An alternative approach is to log-transform lamb weight and then do a standard linear regression analysis. We then use a log-normal distribution. Try out this approach.

Compare the parameter estimates with the ones for the Gamma and inverse Gaussian distributions and comment.

It is also possible to compare the log-Gaussian model with the two previous models using the AIC criterion. Be aware that the response variable is on another scale in the new model. The (log-)likelihood values are therefore not directly comparable.

(e) Let y be our initial response variable and $z = \log(y)$. Show that

$$f_y(y) = f_z(\log(y))/y$$

where f_y is the probability for y while f_z is the probability for z.

Use this to compute the AIC for the log-linear model and compare with the previous models.

Further on we will use the Gamma regression

(f) Discuss why a model without `NAO` as covariate is an improved model.

(g) Examine the final model by

- plotting the residuals
- plotting $\hat{\mu}$ against empirical variance (hint: Look at `car.R` from the lectures).
- goodness-of-fit test

From the general glm-theory we know that $\hat{\boldsymbol{\beta}} \approx N(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is the inverse Fisher information matrix (using the estimated parameter values). If `fit` is you are fitted model (from the glm-routine), then $\boldsymbol{\Sigma}$ is available from the command

```
Sigma = summary(fit)$cov.scaled
```

(h) Assume that $\boldsymbol{x}$ is a vector of covariates and that we are interested in a confidence interval for $\eta = \boldsymbol{x}^T\boldsymbol{\beta}$.

Show that $\hat{\eta} = \boldsymbol{x}^T\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{x}^T\boldsymbol{\beta}, \boldsymbol{x}^T\boldsymbol{\Sigma}\boldsymbol{x})$

Use this to make a 95% confidence interval for η.

Now assume that `aar`=1995, `ageewe`=5, `kjoenn`="m", `burdH`=1 and `alderlam`=150. Compute the confidence interval for η in this case.

(i) Sample $M = 10000$ random variables $\eta_1^*, ..., \eta_M^*$ from $N(\boldsymbol{x}^T\hat{\boldsymbol{\beta}}, \boldsymbol{x}^T\boldsymbol{\Sigma}\boldsymbol{x})$. Show that the empirical 0.025 and 0.975 quantiles from the simulated variables are approximately the same as the CI from the previous task. Use this to make a 95% CI for $\mu = \exp(\eta)$.

(j) The dispersion parameter is estimated by

$$\hat{\phi} = X^2/(n-q)$$

where X^2 is the value of Pearson's test-statistic and q is the number of regression parameters. We also have that $X^2 \approx \phi\chi^2_{n-q}$. Use this to make a 95% confidence interval for ϕ.

(k) Assume that we now are interested in making a 95% *prediction interval* for Y with $\boldsymbol{x}$ given as in the previous subtask. In addition to the uncertainty in $\boldsymbol{\beta}$, the uncertainty in Y will also be included. Consider the following procedure:

(i) For every η_m^*, compute $\mu_m^* = \exp(\eta_m^*)$.

(ii) Simulate $\phi_m^* = \hat{\phi} * (n-q)/\chi^2$ where χ^2 is a variable sampled from χ^2_{n-q}.

(*iii*) Simulate $y_m^*, m = 1, ..., M$ from the Gamma distribution with mean μ_m^* and dispersion parameter ϕ_m^*.

(*iv*) Compute the prediction interval by determine the empirical 0.025 and 0.975 quantile from the sample $y_1^*, ..., y_M^*$.

Perform this procedure to determine a 95% prediction interval for Y given the same $\boldsymbol{x}$ as before.

Hint: The command `rgamma(M,shape=1/phi,scale=mu*phi)` generates `M` variables from the Gamma distribution with the correct mean and dispersion parameters (note that `mu` and `phi` in this case can be vectors).

(*l*) Suggest alternative routines to determine the prediction interval for Y.

Exercise 21

Assume the *random intercept* model where

$$\begin{aligned} Y_{ij} =& \alpha + b_i + \boldsymbol{x}_i^T\boldsymbol{\beta} + \varepsilon_{ij} \qquad\qquad i = 1, ..., N, j = 1, ..., n \\ b_i \sim& N(0, d^2) \\ \varepsilon_{ij} \sim& N(0, \sigma^2) \end{aligned}$$

and the stochastic parts are independent of each other. In this exercise we will look at how the predictions of b_i can be computed. This will be done on the assumption that the parameters are known.

(*a*) Calculate mean and variance for Y_{ij}. Find the covariance between Y_{ij} and Y_{kl}.

(*b*) Show that the distribution for b_i given all other data depends only on data from group i, i.e.

$$p(b_i|\{y_{kj}, k = 1, ..., N, j = 1, ..., n\}) = p(b_i|\{y_{ij}, j = 1, ..., n\}).$$

(*c*) Find the distribution of $\overline{Y}_i = \frac{1}{n}\sum_{j=1}^n Y_{ij}$ given b_i.

Show that $p(b_i|\{y_{ij}, j = 1, ..., n\}) = p(b_i|\overline{y}_i)$, i.e. given the mean, the individual observations will not add any further information on b_i.

(*d*) Find the distribution $p(b_i|\overline{y}_i)$.

Use this to determine

$$\hat{b}_i = E[b_i|\overline{y}_i]$$

Also determine $\text{Var}[b_i|\overline{y}_i]$.

Comment on the results.

Exercise 22

The dataset `Orthodont` is available in R from the package `nlme`. Use the command l`library(nlme)`

The dataset is grouped after the variable `Subject`. The response variable is `distance` with the two covariates `age` and `Sex`. Use `help(Orthodont)` for further info.

In this exercise we will look at how linear mixed models can be used to analyze this dataset.

(a) First transform `Subject` to the values 1-27 using the command

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

(b) Do some exploratory analysis to become familiar with the data set.

(c) Plot `distance` against `age`. For every group, add a regression line with `distance` as response and `age` as covariate.

Comment on the results.

(d) Type in 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

(e) What are the estimates of d^2 and σ^2?

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

Exercise 23

The dataset `petrol` is available from the package `MASS`. Load the dataset and use `help(petrol)` for further info. We will be interested in modeling the response `Y`

The variable `ID` is a group variable we can use to include mixed effects.

Based on model selection approach, chapter 5 in Zuur et al, find the best possible model for this data.

For the final model, try out different routines to see if the model is reasonable.

Exercise 24

In this exercise we will take a closer look on REML estimation in connection with linear regression.

Assume $\boldsymbol{Y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{V}_0)$, i.e. we have extracted σ^2 from the covariance matrix. Our interest will be to estimate σ^2.

We will assume that $\boldsymbol{A}$ is a $n \times (n-p)$ matrix such that $\boldsymbol{A}^T\boldsymbol{X} = \boldsymbol{0}$ with full rank $n-p$ (where we assume that $p < n$, p is the dimension of $\boldsymbol{\beta}$ and n is the dimension of $\boldsymbol{Y}$).

(a) Show that $\boldsymbol{A}^T\boldsymbol{Y}$ has a distribution not depending on $\boldsymbol{\beta}$ and specify this distribution.

(b) Show that

$$\log f(\boldsymbol{A}^T\boldsymbol{Y}) = \text{Const} - \tfrac{n-p}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\boldsymbol{Y}^T\boldsymbol{A}[\boldsymbol{A}^T\boldsymbol{V}_0\boldsymbol{A}]^{-1}\boldsymbol{A}^T\boldsymbol{Y}$$

and use this to show that the ML estimate for σ^2 based on $\boldsymbol{A}^T\boldsymbol{Y}$ is

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

(c) Show that

$$\boldsymbol{Y}^T\boldsymbol{A}[\boldsymbol{A}^T\boldsymbol{V}_0\boldsymbol{A}]^{-1}\boldsymbol{A}^T\boldsymbol{Y} = \text{tr}[[\boldsymbol{A}^T\boldsymbol{V}_0\boldsymbol{A}]^{-1}\boldsymbol{A}^T\boldsymbol{Y}[\boldsymbol{A}^T\boldsymbol{Y}]^T]$$

where tr describes the sum of diagonal elements (trace).

Hint: For the matrices $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ with dimension $r \times s$ and $s \times r$, we have that $\text{tr}[\boldsymbol{M}_1\boldsymbol{M}_2] = \text{tr}[\boldsymbol{M}_2\boldsymbol{M}_1]$.

(d) Show that $E[\boldsymbol{A}^T\boldsymbol{Y}[\boldsymbol{A}^T\boldsymbol{Y}]^T] = \sigma^2\boldsymbol{A}^T\boldsymbol{V}_0\boldsymbol{A}$ and use this to shoe that $\hat{\sigma}^2_{\text{REML}}$ is unbiased.

Hint: Note that the mean and the trace-operation can change order (since trace is a linear operator). On the last part it is important to be sure of the different dimensions on the matrices involved.

(e) Assume that $\widetilde{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{B}$ where $\boldsymbol{B}$ is a non-singular matrix with dimension $(n-p)\times(n-p)$. Show that the REML estimate based on $\widetilde{\boldsymbol{A}}$ are identical with the REML estimate based on $\boldsymbol{A}$.

(f) Assume that $\boldsymbol{X} = \mathbf{1}_N$, i.e. a column vector containing only ones. Further let

$$\boldsymbol{A}^T = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & -1 \\ 0 & 1 & 0 & \cdots & 0 & -1 \\ 0 & 0 & 1 & \cdots & 0 & -1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix}$$

Assume $\boldsymbol{V}_0 = \boldsymbol{I}$.

Determine $\hat{\sigma}^2_{\text{REML}}$ in this case and show that we end up with the common unbiased estimator for σ^2.

Hint: We have that $\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{I} + \mathbf{1}\mathbf{1}^T$. If $\boldsymbol{M} = \boldsymbol{I} + \boldsymbol{v}\boldsymbol{v}^T$ where $\boldsymbol{v}$ is a vector, then $\boldsymbol{M}^{-1} = \boldsymbol{I} - k\boldsymbol{v}\boldsymbol{v}^T$ for suitable choice of k.

Exercise 25 (AR and MA models)

(a) Assume we have a *moving average* model of order 1, MA(1),

$$\varepsilon_s = \theta_1\eta_{s-1} + \eta_s$$

where $\eta_s \overset{iid}{\sim} N(0,\sigma^2)$. Show that ε_s and ε_t are independent for $|s-t| > 1$.

Show that $\varepsilon_s \sim N(0,\tau^2)$ for all s and suitable choice of τ^2.

What is the correlation between ε_s and ε_{s+1}?

(b) Now assume a more general MA(q) model

$$\varepsilon_s = \theta_1\eta_{s-1} + \theta_2\eta_{s-2} + \cdots + \theta_q\eta_{s-q} + \eta_s$$

Show that ε_s and ε_t are independent for $|s-t| > q$.

Show that $\varepsilon_s \sim N(0,\tau^2)$ for all s and suitable choice of τ^2.

What is the correlation between ε_s and ε_{s+v} for $v = 1,...,q$?

(c) Now assume the AR(1) model

$$\varepsilon_s = \phi\varepsilon_{s-1} + \eta_s,$$

where $\eta_s \overset{iid}{\sim} N(0,\sigma^2)$. Show that if $\varepsilon_1 \sim N(0,\tau^2)$ where $\tau^2 = \sigma^2/(1-\phi^2)$, then $\varepsilon_s \sim N(0,\tau^2)$ for all $s > 1$.

Estimate $\text{cor}[\varepsilon_s,\varepsilon_t]$ for all s,t.

Exercise 26
Consider the AR(1) model

$$\varepsilon_s = \phi\varepsilon_{s-1} + \eta_s, \quad \eta_s \overset{iid}{\sim} N(0, \sigma^2)$$

where we assume $\varepsilon_1 \sim N(0, \tau^2)$ and $\tau^2 = \sigma^2/(1-\phi^2)$.

(a) Show that we also can write

$$\text{cor}[\varepsilon_s, \varepsilon_t] = \exp\{-|t-s|/d\}$$

where $d = -1/\log(\phi)$.

The advantage of this formulation is that we are able to generalize the model to situations where the time between time points may differ. Such a correlation structure is described as a *exponential correlation function* and d is often called the *range* parameter.

(b) The dataset `Hawaii` is available from the package `STK3100`. Load the dataset and use the commands below

```
Hawaii$Birds <- sqrt(Hawaii$Moorhen.Kauai)
M0 = gls(Birds ~ Rainfall+Year,na.action=na.omit,data=Hawaii)
M1 = gls(Birds ~ Rainfall + Year, na.action = na.omit,
      correlation = corAR1(form =~ Year), data = Hawaii)
Hawaii2 = Hawaii[!is.na(Hawaii$Birds),]
M2 = gls(Birds ~ Rainfall + Year, data = Hawaii2,
      correlation = corExp(form =~ Year))
```

The first commands produce the AR(1) model as shown in class. The last two commands use the exponential correlation function. Note that we do not need to have one row per year even when `Hawaii2` have some years missing. It is therefore not necessary to use the extra command `na.action = na.omit`.

Show that $\hat{d} = -1/\log(\hat{\phi})$

We will now look at the dataset `spruce`, also available from the package `STK3100`. The dataset shows growth of different trees and includes the following covariates:

- `Tree`, index factor
- `days`, numeric variable indicating the number of days since the start of the experiment
- `logSize`, a numeric value that gives the estimated logarithm of the volume of the tree trunk.

- `plot`, a factor that identifies the piece of land where the tree is located.

We will start by analyzing a single tree, which we select with the command

```
Spruce1 = Spruce[Spruce$Tree=="01T18",]
```

(c) Plot `logSize` against `days`. Comment.

Since the time varies will the exponential correlation function be suitable.

(d) Try out a model with `days` as a linear fixed effect and independence between the error terms. Compare it with a model where you use the exponential correlation function.

Which model seems to be best suited in this case?

What does it mean to the estimates, standard error and p-values that we include time dependencies?

(e) We will now include all the trees in the analysis. It is then possible to include `plot` as a covariate. Try out the commands

```
M0 = gls(logSize ~ days+plot,Spruce)
cexp = corExp(form =~days|Tree,fixed=FALSE)
M1 = gls(logSize ~ days+plot,Spruce,correlation = cexp)
```

The code `form =~days|Tree` means that the exponential correlation function will be used on every single tree. The trees indicate the group structure and we want independence between the groups as before.

Which model seems to be best suited in this case?

What does it mean to the estimates, standard error and p-values that we include time dependencies in this case?

Within the two models for the correlation structure, perform model selection for the fixed effects.Will the two models for correlation structure have influence on the model selection for the fixed effects?

Hint: Remember that `gls` uses REML as default for estimation.

Exercise 27

We will in this exercise look at the situation where

$$y_{ij} \sim N(\mu_i, \sigma^2), i = 1, ..., n, j = 1, ..., m$$

and all the observations are independent.

We will be interested in the estimation of σ^2 and look at how the estimate for σ^2 behave when n increases.

Note that the number of parameters increases with n in this case.

We will in the first part assume that $m = 2$.

(*a*) Find the maximum likelihood estimates for $\mu_i, i = 1, ..., n$ and σ^2.

(*b*) Find $E[\hat{\sigma}^2_{ML}]$ and show that $\lim_{n\to\infty} \hat{\sigma}^2_{ML} \neq \sigma^2$, that is, the ML estimate are not consistent.

Discuss how this is related to the general ML-theory (i.e. which assumptions are not met in this situation?)

(*c*) Now let $z_i = y_{i1} - y_{i2}$. Show that the ML-estimate σ^2 based on $z_1, ..., z_n$ is consistent and unbiased.

(*d*) Show that the estimation in the previous exercise is in fact REML estimation.

(*e*) Consider a general m and find the ML-estimates for $\mu_1, ..., \mu_n$ and σ^2 based on $\{y_{ij}, i = 1, ..., n, j = 1, ..., m\}$.

(*f*) Determine $E[\hat{\sigma}^2_{ML}]$. What is required for $\hat{\sigma}^2_{ML}$ to be a consistent estimator?

Exercise 28

We will in this exercise look at a dataset from Steele (1998), *the Australian Data and Story Library* (OzDASL). An experiment was conducted to study the effect of surface and vision on balance. The balance of individuals were observed for two different surfaces (plain and a foam-like surface) and for different types of visual skills (eyes open, closed or partially limited). Balance was rated on a scale of 1 to 4, but we will here look at a binary response, where 1 corresponds to 1 while 0 corresponds to 2,3,4. For each individual gender, age, height and weight were registered. Each individual was tested twice for each combination of different surface and vision, a total of 12 observations. The dataset `ctsib` is available from the package `STK3100`. `stable` is our binary response (`CTSIB` is the initial qualitative measure).

(*a*) Start with an ordinary glm-fit with `Sex,Age,Height,Weight,Surface` and `Vision` as covariates. Explain why we can not trust the p-values.

(*b*) Include `Subject` as a factorial covariate.

What difficulties do you encounter for this model-fit (look at the summary).

Compare with the previous model using the `anova` function.

What is the disadvantage of specifying `Subject` as a fixed effect?

(*c*) Try to include `Subject` as a random effect.

Discuss the advantages of this kind of model.

(*d*) Try out a model selection strategy to determine the important covariates.

What model do you end up with?

What effect does it have on the random effects when some fixed effects are removed?

Hint: Use the `lmer` routine since `glmmPQL` is not suitable for model comparison. You can use the `anova` routine to compare several models.

Exercise 29 (ANOVA and LMM models)
Consider a single factor ANOVA model

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2), \quad i = 1, ..., I, j = 1, ..., J \qquad (*)$$

where for identifiability we impose the constraints $\sum_{i=1}^{I} \alpha_i = 0$. Important quantities when analyzing such models are

$$\begin{aligned} \text{SSTr} &= J \sum_{i=1}^{I} (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 \\ \text{SSE} &= \sum_{i=1}^{I} \sum_{j=1}^{J} (y_{ij} - \bar{y}_{i\cdot})^2 \end{aligned}$$

where $\bar{y}_{i\cdot} = \frac{1}{J} \sum_{j=1}^{J} y_{ij}$ and $\bar{y}_{\cdot\cdot} = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} y_{ij}$.

(a) By looking at textbooks from earlier courses or by your own calculations, find the expectations of SSTr and SSE under model (*).

Explain how these quantities can be used to test the hypothesis H_0 : $\alpha_1 = \alpha_2 = \cdots = \alpha_I = 0$.

(b) An alternative random effects formulation of the model above is

$$Y_{ij} = \mu + A_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2), \quad i = 1, ..., I, j = 1, ..., J \quad (**)$$

where $A_i \overset{iid}{\sim} N(0, \sigma_A^2)$ and all A_i's independent of all ε_{ij}'s.

Show that this is a special case of a linear mixed model (LMM).

(c) Find the expectations of SSTr and SSE under model (**).

Hint: For finding the expectation of SSE, show that this quantity does not depend on the α_i's and argue why you then can use results from (a). For finding the expectation of SSTr, show first that $\overline{Y}_{i\cdot}$ are iid and Gaussian.

(d) Show that

$$\begin{aligned} \hat{\sigma}^2 &= \frac{\text{SSE}}{I(J-1)} \\ \hat{\sigma}_A^2 &= \frac{\text{SSTr}}{J(I-1)} - \frac{\hat{\sigma}^2}{J} \end{aligned}$$

are unbiased estimates for σ^2 and σ_A^2, respectively.

We will now consider a simulation study, where we generate data according to model (**) and explore the behavior of the unbiased estimates as well as the ML estimates (obtained by using `lme`). For this part, there is an **R** script `sim_raneff.R` available from the course home-page which performs the simulations for you.

(e) Discuss the two plots you obtain. In particular, comment on the cases where $\hat{\sigma}_A^2 < 0$. Based on these plots, give an approximate relationship between the unbiased estimates and the ML estimates.

(f) Now modify the script so that $\sigma_A^2 = 0$. How many of the simulations result in that $\hat{\sigma}_A^2 < 0$. Up to 4 digits of precision, what are the values of -2LR for these simulations?

(g) Make a histogram of -2LR for those simulations corresponding to $\hat{\sigma}_A^2 > 0$. Compare this histogram with a χ_1^2 density.

(h) Discuss these results related to the general result that -2LR for testing $H_0 : \sigma_A^2 = 0$ is approximately a mixture of a χ_0^2 and a χ_1^2 distribution (with equal weight on each), where here χ_0^2 is a distribution putting all weight in 0.

Modify the script to include a test on H_0 based on LR. How many times is H_0 rejected?

(i) An alternative to likelihood ratio tests in this case is to use an F test directly through the use of SSE and SSTr. Devore and Beck (2007) state that under H_0,

$$F = \frac{SSTR/(I-1)}{SSE/(I(J-1))}$$

follows an F-distribution with $I-1$ and $I(J-1)$ degrees of freedom.

Include in the script a test based on this approach and compare with the LR test.

Comment on the results.

Remarks: Although the F test is easier to use in this case, such a test will not be possible to use in more general settings such as unbalanced designs and/or nonlinear models. In such cases, likelihood ratio tests needs to be used.

Exercise 30

The `abrasion` data frame has 16 rows and 4 columns and is available from the package `STK3100`. Four materials were fed into a abrasion testing machine and the amount of wear recorded. Four samples could be processed at the same time and the position of these samples may be important. A

Latin square design was used (this detail is not important for us here). A possible model for these data are

$$y_{ijk} = \beta_0 + \alpha_i + \eta_j + \gamma_k + \varepsilon_{ijk}$$

where i is an index for the type of material, j an index for the position and k for the run. Here β_0 and the α_i's are treated as fixed effects while $\eta_j, \gamma_k, \varepsilon_{ijk}$ are random quantities.

(a) Show that the model can be written as a linear mixed model and specify the parameters involved.

(b) This model can be fitted by the `lmer` routine (`lme` apparently do not cover this case, use `library(lme4)`) by the command

```
fit = lmer(wear~material+(1|run)+(1|position),abrasion)
```

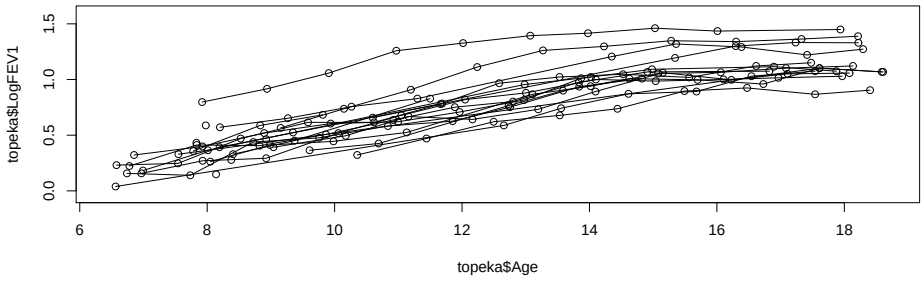
Look at the output from this call and specify the estimates of the parameters involved.

(c) Perform tests to check whether a simpler structure on the random effects can be used.

(d) Also perform a test for checking whether `material` should be part of the model.

Exercise 31 ()
In a larger study to examine the factors that describe lung capacity over time (measured as changes in lung function in the teens), a total of 13,379 children were studied with respect to lung function at different times. We will look at a small subset consisting of 300 girls where changes in lung function was measured by maximal inhalation followed by exhaust pressure as quickly as possible in a closed container. Total volume blown out the first second is registered as FEVi.

The plot below shows log(FEV1) against age for 25 randomly selected girls.



We will consider the following model:

$$
\begin{aligned}
Y_{ij}|\boldsymbol{b}_i &\overset{ind}{\sim} N(\mu_{ij}, \sigma^2) \\
\mu_{ij} &= \beta_0 + \beta_1 \text{Age}_{ij} + \beta_2 \log(\text{Ht}_{ij}) + \beta_3 \text{Age}_{ij} \log(\text{Ht}_{ij}) + b_i \\
b_i &\overset{iid}{\sim} N(\mathbf{0}, d)
\end{aligned}
$$

where $Y = \log(\text{FEV}_1)$, Age_{ij} is age at the measurements, and where Ht_{ij} is related height.

(a) Discuss the advantages of using random effects for the analysis on this dataset.

What is the correlation between two observations from the same individual in this model?

(b) Why is it suitable to include fixed effects in the model early on before the structure of the random effects are determined?

The output underneath shows the results for the fitted model above using REML estimation.

```
Linear mixed model fit by REML
Formula: LogFEV1 ~ Age * log(Ht) + (1 | ID)
   Data: topeka
   AIC   BIC logLik deviance REMLdev
```

```
 -4515 -4482   2264    -4562   -4527
Random effects:
 Groups   Name        Variance  Std.Dev.
 ID       (Intercept) 0.0090710 0.095242
 Residual             0.0040626 0.063738
Number of obs: 1993, groups: ID, 299

Fixed effects:
              Estimate Std. Error t value
(Intercept) -0.1777251  0.0323194  -5.499
Age         -0.0008542  0.0043038  -0.198
log(Ht)      1.9188462  0.0648401  29.593
Age:log(Ht)  0.0453024  0.0075763   5.980
```

(*c*) Explain what we mean by REML estimation in this case. Why is it beneficial to use REML estimation in models with mixed effects?

Determine the estimated correlation between two observations from the same individual in this case.

(*d*) A model where we included a random effect multiplied with $\log(\text{Ht})_{ij}$ gave the following results:

```
Linear mixed model fit by REML
Formula: LogFEV1 ~ Age * log(Ht) + (1 + log(Ht) | ID)
   Data: topeka
   AIC   BIC logLik deviance REMLdev
 -4629 -4584   2322    -4678   -4645
Random effects:
 Groups   Name        Variance  Std.Dev. Corr
 ID       (Intercept) 0.0138480 0.117678
          log(Ht)     0.0784168 0.280030 -0.645
 Residual             0.0033921 0.058242
```

Another model where we also included a random effect multiplied with Age_{ij} gave the following results:

```
Linear mixed model fit by REML
Formula: LogFEV1 ~ Age * log(Ht) + (1 + Age + log(Ht) | ID)
   Data: topeka
   AIC   BIC logLik deviance REMLdev
 -4624 -4563   2323    -4680   -4646
Random effects:
 Groups   Name        Variance   Std.Dev.  Corr
 ID       (Intercept) 1.3935e-02 0.1180468
          Age         2.1296e-05 0.0046147 -0.266
```

```
          log(Ht)     7.6638e-02 0.2768358 -0.517 -0.237
 Residual              3.3631e-03 0.0579925
```

What is the difference in the number of parameters between the various models?

Use this to consider which model is best according to the likelihood ratio principle.

Does your conclusion match with what you get using the AIC/BIC criteria?

(e) We will now look at the fixed effects. Why is it suitable to use REML estimation when we are comparing models with different fixed effects?

(f) The model with two random effects and ML estimation gave the following results for the fixed effects:

```
Fixed effects:
             Estimate Std. Error t value
(Intercept) -0.110366   0.035296  -3.127
Age         -0.011651   0.004626  -2.519
log(Ht)      1.849369   0.069083  26.770
Age:log(Ht)  0.063350   0.008194   7.731
```

Use this to argue why there is no reason to remove any of the fixed effects.

(g) Output from the final model based on REML estimation is given below. Use this to determine $\mu = E[Y]$ for Age $= 9.3415$ and Ht $= 1.20$, estimate related standard deviation and determine a 95% confidence interval for μ.

Hint: Remember the correlation between the β-estimates.

```
Linear mixed model fit by REML
Formula: LogFEV1 ~ Age * log(Ht) + (1 + log(Ht) | ID)
   Data: topeka
   AIC   BIC logLik deviance REMLdev
 -4629 -4584   2322    -4678   -4645
Random effects:
 Groups   Name        Variance  Std.Dev. Corr
 ID       (Intercept) 0.0138480 0.117678
          log(Ht)     0.0784168 0.280030 -0.645
 Residual             0.0033921 0.058242
Number of obs: 1993, groups: ID, 299

Fixed effects:
```

```
             Estimate Std. Error t value
(Intercept) -0.110118   0.035344  -3.116
Age         -0.011694   0.004631  -2.525
log(Ht)      1.849176   0.069173  26.733
Age:log(Ht)  0.063419   0.008204   7.730

Correlation of Fixed Effects:
            (Intr) Age    lg(Ht)
Age         -0.932
log(Ht)     -0.814  0.608
Age:log(Ht)  0.963 -0.967 -0.770
```