1.
```
> d = read.delim("http://dnett.github.io/S510/LeafArea.txt")
> library(lme4)
>
> o = lmer(LeafArea ~ Dose + (1 + Dose | ResearchStation), data = d)
> summary(o)
Linear mixed model fit by REML ['lmerMod']
Formula: LeafArea ~ Dose + (1 + Dose | ResearchStation)
   Data: d

REML criterion at convergence: 1333.905

Random effects:
 Groups          Name        Variance  Std.Dev. Corr
 ResearchStation (Intercept) 1.049e+01 3.238634
                 Dose        5.623e-05 0.007499 0.06
 Residual                    3.949e+00 1.987147
Number of obs: 300, groups: ResearchStation, 15

Fixed effects:
             Estimate Std. Error t value
(Intercept) 13.857667   0.859498   16.12
Dose         0.051900   0.003779   13.73

Correlation of Fixed Effects:
     (Intr)
Dose -0.131
>
> u=ranef(o)$ResearchStation
> b=fixef(o)
> vcov(o)
2 x 2 Matrix of class "dpoMatrix"
              (Intercept)          Dose
(Intercept)  0.7387375451 -0.0004269892
Dose        -0.0004269892  0.0000142787
```

   (a) $\hat{\sigma}_e^2 = 3.949$

   (b)
$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{b}} = \begin{bmatrix} 10.49 & 0.06 * \sqrt{10.49 * 0.00005623} \\ 0.06 * \sqrt{10.49 * 0.00005623} & 0.00005623 \end{bmatrix}$$

   (c)
```
> plot(d$Dose[d$ResearchStation == 7], d$LeafArea[d$ResearchStation == 7],
+      xlab = "Dose", ylab = "Leaf Area")
> lines(c(0,100), b[1] + (b[2]) * c(0,100))
```
   See actual figure after part (f).

   (d)
```
> #eBLUP of intercept for research station 7.
```

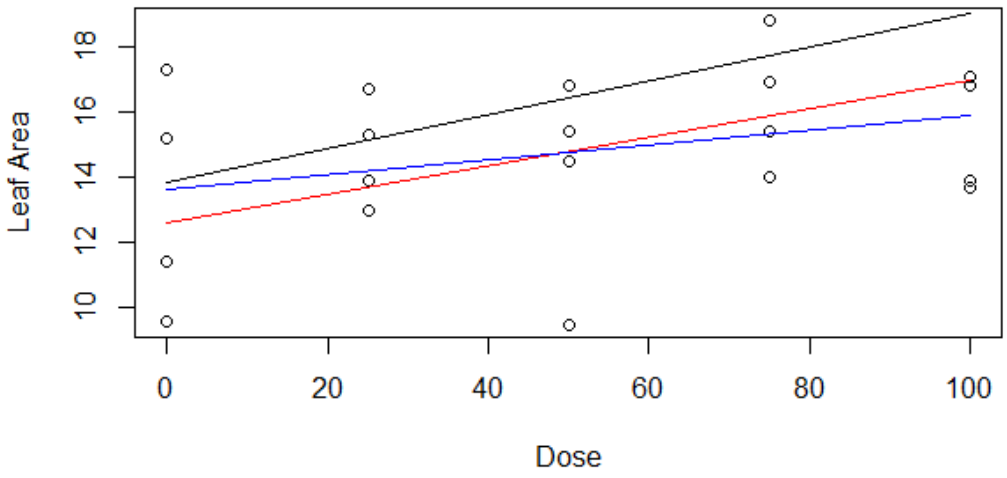
```
>
> b[1] + u[7,1]
(Intercept)
   12.59071
>
> #eBLUP of slope for research station 7.
>
> b[2] + u[7,2]
      Dose
0.04378323
```

$12.59071 + 0.04378323x$

(e)
```
> #Simple linear regression for research station 7
>
> o7 = lm(d$LeafArea[d$ResearchStation == 7] ~ d$Dose[d$ResearchStation == 7])
> b7 = coef(o7)
> b7
                    (Intercept) d$Dose[d$ResearchStation == 7]
                        13.6500                         0.0222
```

$13.6500 + 0.0222x$

(f)
```
> lines(c(0,100), b[1] + u[7,1] + (b[2] + u[7,2]) * c(0,100), col = "red")
> lines(c(0,100), b7[1] + (b7[2]) * c(0,100), col = "blue")
```



Note that, relative to the blue simple linear regression fit, the red eBLUP regression line is rotated towards the population regression line in black. The slopes of the red and black lines are similar because the small estimated variance for the slope random effects means there will be strong shrinkage of the eBLUP predicted slope towards the slope of the black line.

(g) To get the correct likelihood ratio statistic, we need to be sure to ask R to use ML rather than REML to estimate parameters in this case. It doesn't make sense to compare REML likelihoods here because the two models have different models for the mean of $\boldsymbol{y}$.

```
> null.model = lmer(LeafArea ~ 1 + (1 + Dose | ResearchStation),
+                   data = d, REML = F)
> alt.model = lmer(LeafArea ~ Dose + (1 + Dose | ResearchStation),
+                  data = d, REML = F)
>
> anova(null.model, alt.model, test = "Chisq")
Data: d
Models:
null.model: LeafArea ~ 1 + (1 + Dose | ResearchStation)
alt.model: LeafArea ~ Dose + (1 + Dose | ResearchStation)
           Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
null.model  5 1376.1 1394.6 -683.06   1366.1
alt.model   6 1338.0 1360.3 -663.02   1326.0 40.078      1   2.44e-10 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Based on the code and output above, the likelihood ratio statistic is 40.078.

(h)
```
> ofull = lmer(LeafArea ~ Dose + (1 + Dose | ResearchStation), data=d, REML=F)
> AIC(ofull)
[1] 1338.042
```

See below for AIC computed using REML.

(i)
```
> o1slope = lmer(LeafArea ~ Dose + (1 | ResearchStation), data = d, REML = F)
> AIC(o1slope)
[1] 1334.575
```

See below for AIC computed using REML.

(j)
```
> o1int1slope = lm(LeafArea ~ Dose, data = d)
> AIC(o1int1slope)
[1] 1650.107
```

Verification that the above is indeed the likelihood version of AIC – and thus comparable to the calculations in parts (h) and (i) – comes from the following calculations.

$$
\begin{aligned}
\ell(\hat{\boldsymbol{\theta}}) &= -\frac{1}{2}\log|\hat{\boldsymbol{\Sigma}}| - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}) - \frac{n}{2}\log(2\pi) \\
&= -\frac{1}{2}\log\left|\frac{\text{SSE}}{n}\boldsymbol{I}\right| - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})'\left(\frac{\text{SSE}}{n}\boldsymbol{I}\right)^{-1}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}) - \frac{n}{2}\log(2\pi) \\
&= -\frac{n}{2}\log\left(\frac{\text{SSE}}{n}\right) - \frac{1}{2\frac{\text{SSE}}{n}}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}) - \frac{n}{2}\log(2\pi) \\
&= -\frac{n}{2}\log\left(\frac{\text{SSE}}{n}\right) - \frac{n}{2\text{SSE}}\text{SSE} - \frac{n}{2}\log(2\pi) \\
&= -\frac{n}{2}\log\left(\frac{\text{SSE}}{n}\right) - \frac{n}{2} - \frac{n}{2}\log(2\pi)
\end{aligned}
$$

```
> #AIC for likelihood following R convention for AIC
>
```

```
> SSE = sum(residuals(o1int1slope)^2)
>
> logMLlike = -0.5 * 300 * log(SSE / 300) - 0.5 * 300  - (300 / 2) * log(2*pi)
>
> -2 * logMLlike + 2 * 3
[1] 1650.107
>
> #Alternatively, using the R function logLik
>
> -2 * logLik(o1int1slope) + 2 * 3
[1] 1650.107
```

The answers for AIC computed using REML are as follows.

```
> ofull = lmer(LeafArea ~ Dose + (1 + Dose | ResearchStation), data = d)
> AIC(ofull)
[1] 1345.905
>
> o1slope = lmer(LeafArea ~ Dose + (1 | ResearchStation), data = d)
> AIC(o1slope)
[1] 1342.693
>
> #According to my reading of R help files, the following should give AIC
> #for REML with the lm function.
>
> o1int1slope = lm(LeafArea ~ Dose, data = d)
>
> - 2 * logLik(o1int1slope, REML = T) + 2 * 3
'log Lik.' 1659.678 (df=3)
```

(k) Whether we use ML or REML to get AIC, the model from part (i) is preferred because it has the lowest AIC.

2. Model (1) from problem 1 can be written as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u}+\boldsymbol{\epsilon}$. Define $\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{Z}, \boldsymbol{u}, \boldsymbol{G} = Var(\boldsymbol{u})$ and $\boldsymbol{R} = Var(\boldsymbol{\epsilon})$:

$$\boldsymbol{X} = \mathbf{1}_{15\times 1} \otimes \begin{bmatrix} 1 & 0 \\ 1 & 25 \\ 1 & 50 \\ 1 & 75 \\ 1 & 100 \end{bmatrix} \otimes \mathbf{1}_{4\times 1}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix},$$

$$\boldsymbol{Z} = \boldsymbol{I}_{15} \otimes \begin{bmatrix} 1 & 0 \\ 1 & 25 \\ 1 & 50 \\ 1 & 75 \\ 1 & 100 \end{bmatrix} \otimes \mathbf{1}_{4\times 1}, \quad \boldsymbol{u} = (b_{11}, b_{21}, \cdots, b_{1,15}, b_{2,15})',$$

$\boldsymbol{G} = \boldsymbol{I}_{15} \otimes \boldsymbol{\Sigma_b}$ where $\boldsymbol{\Sigma_b}$ is a 2×2 variance matrix for $(b_{1i}, b_{2i})'$ for any $i = 1, \cdots, 15$, and $\boldsymbol{R} = \sigma_e^2 \cdot \boldsymbol{I}_{300}$.

3. Several models could be considered for this dataset. The analysis presented here is for the additive model that includes age as a quantitative variable and sex as a factor. In this model, the logit of the survival probability is a linear function of age, with a different intercept for each sex.

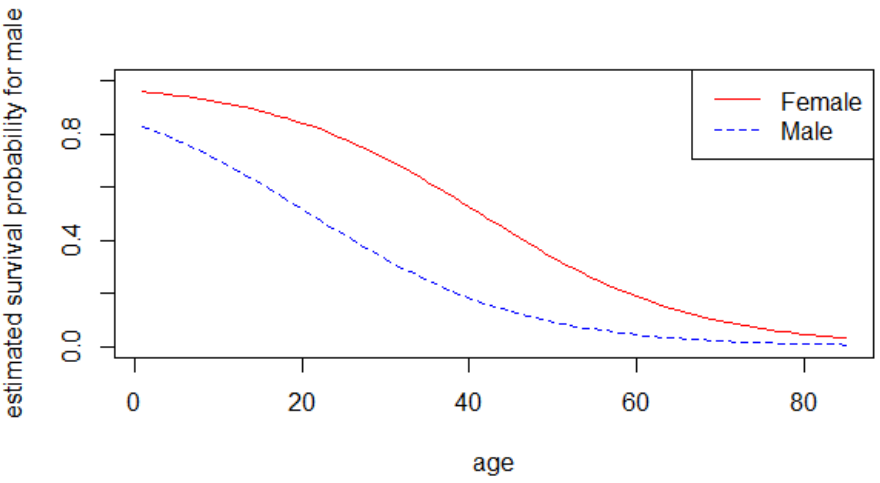
   It is not a good idea to treat age as a factor because there are 21 different ages among the 45 observations and thus few observations per parameter if age is treated categorically. Furthermore, there are several ages with only one observation, which leads to numerical problems with model fitting in which logits need to be $\pm\infty$ to obtain success probability estimates of 0 or 1. For example, there is one person age 21 in the dataset who survived. If age is a factor, the effect estimate for age 21 should be ∞ so that the estimated survival probability is 1 for people age 21 because a survival probability of 1 for people age 21 will maximize the likelihood. The numerical algorithm that attempts to maximize the log likelihood can essentially march on forever and continually increase the likelihood by raising the effect estimate for age 21 towards infinity. It is better and more meaningful to treat age quantitatively in this case. This allows for a smooth effect of age on the survival probability. A linear effect for age in the logit is not necessarily the true model, but it does give some useful information and interpretations. Higher order effects (such as quadratic) could be considered, but in this case linear seems to fit better than quadratic based on statistical testing and information criteria (AIC, BIC) analysis.

   A model that allows for interaction between age and sex has some merit. Although the age-by-sex interaction is not significant at the 0.05 level (p-value = 0.0865), the interaction model is favored by AIC and BIC. For simplicity, these solutions report the results for the additive model only.

```
> donner=read.delim("http://dnett.github.io/S510/Donner.txt")
> attach(donner)
> glm=glm(status~age+sex,family=binomial(link=logit))
> summary(glm)

Call:
glm(formula = status ~ age + sex, family = binomial(link = logit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7445  -1.0441  -0.3029   0.8877   2.0472

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.23041    1.38686   2.329   0.0198 *
age         -0.07820    0.03728  -2.097   0.0359 *
sexMALE     -1.59729    0.75547  -2.114   0.0345 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 61.827  on 44  degrees of freedom
```

```
Residual deviance: 51.256  on 42  degrees of freedom
AIC: 57.256
```

From the above results we can observe that expected survival probability is higher for women. Age also had a significant association with survival probability. With age increasing, the expected survival probability will be decreased. The estimated probability plot also shows the same pattern as follows:




```
> x=1:85
> b=coef(glm)
> #plot estimated survival prob.
> plot(x,1/(1+exp(-b[1]-b[2]*x)),ylim=c(0,1),type="l",col="red",xlab="age",ylab
+      ="estimated survival probability for male",lty=1)
> lines(x,1/(1+exp(-b[1]-b[3]-b[2]*x)),col="blue",lty=2)
> leg=c("Female","Male")
> legend("topright",legend=leg,col=c('red','blue'),lty=1:2)
```

The output below shows that a one year increase in age is associated with a 7.52% decrease in the odds of survival for people with the same sex. For a given sex, we are 95% confident that one year increase in age is associated with a multiplicative change in the odds of survival between (0.850, 0.986). [To understand this better, let o_{x+1} be the odds of survival at age $x+1$, and let o_x be the odds of survival at age x. Then from our course notes, we know that $o_{x+1} = \exp(\hat{\beta}_{age})o_x$. Note that $\exp(\hat{\beta}_{age}) \approx \exp(-0.07820407) \approx 0.9247757$. This means the odds at age $x+1$ are estimated to be about 92.5% times the odds at age x. This is a decrease in the odds of about 100-92.5=7.5%.]

```
> 1-exp(b[2])
       age
0.07522431
> ci=confint(glm)
Waiting for profiling to be done...
```

```
> exp(ci[2,])
    2.5 %    97.5 %
0.8500691 0.9860327
```

For people of the same age, a male's odds of survival are 79.8% lower than the female's odds of survival. For a given age, we are 95% confident that being a male is associated with a multiplicative change in the odds of survival between (0.0396, 0.8227).

```
> 1-exp(b[3])
  sexMALE
0.7975563
> exp(ci[3,])
     2.5 %     97.5 %
0.03961013 0.82275075
```

4. Let y_{ij} be the response value for the jth experimental unit treated with treatment i (where $i = 1, 2, 3$ correspond to treatments A, B, and C, respectively). The model the researchers fit to the data is

$$y_{ij} = \mu_i + e_{ij}, \text{ where the } e_{ij} \text{ terms are assumed to be iid } N(0, \sigma^2). \quad (1)$$

This model (1) implies that each y_{ij} is normally distributed. This differs from reality because each y_{ij} is actually binomially distributed. That may not be a big problem, though, because binomial distributions can be reasonably well approximated by normal distributions. The model also assumes independence, but this seems reasonable based on the information given. The most serious problem in this case is the constant variance assumption that says every error (and thus every response) has the same variance σ^2; i.e., $\text{Var}(y_{ij}) = \sigma^2$ for all i and j. Recall that the variance for a single Binomial(m, p) observation is $mp(1-p)$. For this experiment, we have $\text{Var}(y_{1j}) = \text{Var}(y_{2j}) = 5$ and $\text{Var}(y_{3j}) = 0.95$ as indicated in Table 1.

Table 1: True variance of an observation (one experimental unit) for each treatment.

Treatment	p	Variance of an Observation
A	0.5	$20(0.5)(1-0.5) = 5.00$
B	0.5	$20(0.5)(1-0.5) = 5.00$
C	0.95	$20(0.95)(1-0.95) = 0.95$

What will happen when the researcher uses the test for "trtB" provided by R in the output from the fit of model (1)? The test statistic is

$$\frac{\bar{y}_{1\cdot} - \bar{y}_{2\cdot}}{\sqrt{\text{MSE}(\frac{1}{10} + \frac{1}{10})}}.$$

Ideally, the term under the square root sign would be an estimate of the variance of the numerator. However, the variance of the numerator is

$$\text{Var}(\bar{y}_{1\cdot} - \bar{y}_{2\cdot}) = \text{Var}(\bar{y}_{1\cdot}) + \text{Var}(\bar{y}_{2\cdot}) = \frac{5}{10} + \frac{5}{10} = 1,$$

while the expected value of the term under the square root sign is

$$\frac{(10-1)*5 + (10-1)*5 + (50-1)*0.95}{70-3}\left(\frac{1}{10} + \frac{1}{10}\right) \approx 0.408.$$

Thus, the denominator of the statistic will tend to be considerably smaller than it should be, which will make values far from zero more likely than they should be. Because the null hypothesis the researcher is testing is true, the probability of a type I error will be inflated. For example, the probability of a p-value less than 0.05 will be higher than 0.05 even though the null hypothesis is true.