hw12

Yonghyun Kwon

4/28/2020

Problem 1

```
d = read.delim("http://dnett.github.io/S510/LeafArea.txt")
o = lmer(LeafArea ~ Dose + (1 + Dose | ResearchStation), data = d, REML = TRUE,
         control = lmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 100000)))
```

(a)

REML estimate of σ_e^2 is

```
sigma(o)^2
```

```
## [1] 3.948756
```

(b)

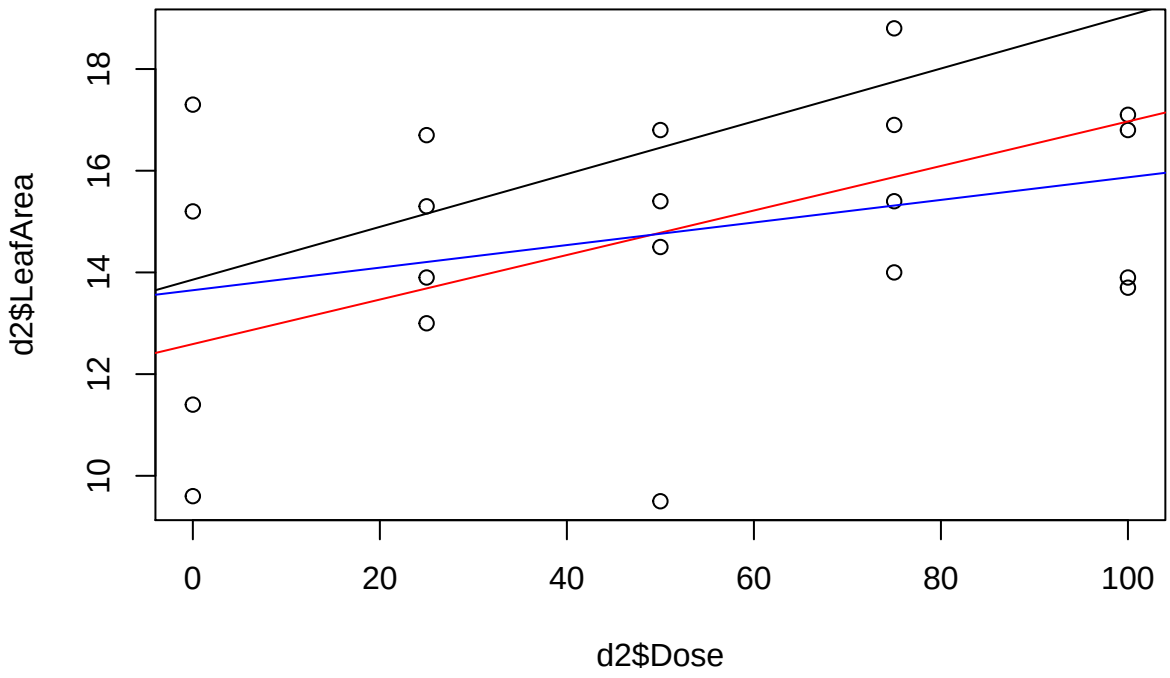
REML estimate of Σ_b is

```
VarCorrdat <- as.data.frame(VarCorr(o))$vcov
Sigma <- matrix(c(VarCorrdat[1], VarCorrdat[3],
                  VarCorrdat[3], VarCorrdat[2]), nr = 2)
round(Sigma, 6)
```

```
##           [,1]     [,2]
## [1,] 10.488700 0.001493
## [2,]  0.001493 0.000056
```

(c) and (f)

```
d2 <- filter(d, d$ResearchStation == 7)
plot(d2$Dose, d2$LeafArea)
beta <- fixef(o)
b <- ranef(o)$ResearchStation
abline(beta)
beta0 <- unlist(beta + b[7,])
o2 = lm(LeafArea ~ Dose, data = d2)
beta2 <- coef(o2)
abline(beta0, col = "red")
abline(beta2, col = "blue")
```



(d)

```
sprintf("%gx +%g", beta0[1], beta0[2])
```

```
## [1] "12.5907x +0.0437833"
```

(e)

```
sprintf("%gx +%g", beta2[1], beta2[2])
```

```
## [1] "13.65x +0.0222"
```

(g)

To compute the likelihiood ratio statistic for testing $H_o : \beta_2 = 0$, use *anova* function in R

```
o3 <- lmer(LeafArea ~ 1 + (1 + Dose | ResearchStation), data = d, REML = FALSE,
           control = lmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 100000)))
anova(o, o3)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: d
## Models:
```

```
## o3: LeafArea ~ 1 + (1 + Dose | ResearchStation)
## o: LeafArea ~ Dose + (1 + Dose | ResearchStation)
##    Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## o3  5 1376.1 1394.6 -683.06   1366.1
## o   6 1338.0 1360.3 -663.02   1326.0 40.086      1   2.43e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that R automatically refit the models with ML because REML-based log-likelihood is no longer meaningful when testing different mean models. According to the table above, the likelihood ratio statistics is 40.086.

(h)

```
AIC(o)
```

```
## [1] 1345.905
```

(i)

```
AIC(lmer(LeafArea ~ Dose + (1 | ResearchStation), data = d, REML = TRUE,
         control = lmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 100000))))
```

```
## [1] 1342.693
```

(j)

```
AIC(logLik(lm(LeafArea ~ Dose, data = d), REML = TRUE))
```

```
## [1] 1659.678
```

##(k) Since AIC for the second model is the smallest, model considered in part (i) is preferred.

Problem 2

X, β, Z, u, G, R are

$$X = \mathbb{1}_{15} \otimes \begin{pmatrix} \mathbb{1}_5 & x \end{pmatrix} \otimes \mathbb{1}_4,\ \beta = (\beta_1 \beta_2)^T$$
$$Z = I_{15} \otimes \begin{pmatrix} \mathbb{1}_5 & x \end{pmatrix} \otimes \mathbb{1}_4,\ u = (b_{11}, b_{21}, b_{12}, b_{22}, \cdots, b_{115}, b_{215})^T$$
$$G = I_{15} \otimes \Sigma_b$$
$$R = \sigma_e^2 I_{300}$$

where $x = (x_1, x_2, x_3, x_4, x_5)^T$

Problem 3

```
Donner = read.delim("https://dnett.github.io/S510/Donner.txt")
Donner$status<-ifelse(Donner$status=="SURVIVED", 1, 0)
```

To explain how age and sex are associated with the probability of survival, we construct the following model

$$survival_i \sim Bernoulli(\pi_i) \text{logit}(\pi_i) = x_i^T \beta_i$$

where $x_i = (1, age_i, I(sex_i = Male))^T$. In R, the model is parametrized and expressed as

```
model <- glm(status ~ age + sex, family = binomial(link = "logit"), data = Donner)
summary(model)
```

```
##
## Call:
## glm(formula = status ~ age + sex, family = binomial(link = "logit"),
##     data = Donner)
##
## 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
##
## Number of Fisher Scoring iterations: 4
```

To check that the model is reliable, observe that the deviance and degree of freedom for residual is

```
deviance(model)
```

```
## [1] 51.25628
```

```
df.residual(model)
```

```
## [1] 42
```

```
deviance(model)/df.residual(model)
```

```
## [1] 1.220388
```

Since deviance is not significantly larger than the residual, there is no overdispersion. This also implies that there is no significant lack of fit because

```
1-pchisq(deviance(model), df.residual(model))
```

```
## [1] 0.1549077
```

The p-value for lack of fit test is greater than 0.05.

Now, from this model, consider an appropriate test

```r
summary(model)
```

```
##
## Call:
## glm(formula = status ~ age + sex, family = binomial(link = "logit"),
##     data = Donner)
##
## 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
##
## Number of Fisher Scoring iterations: 4
```

For example, the following test can be used to test whether sex main effect is significant.

$$H_0 : sex_{MALE} = 0,\ H_1 : \text{not } H_0$$

Since p-value for *sexMALE* in the above summary table is 0.0345, which is smaller than 0.05, sex main effect is significant with significance level 0.05.

Similarly, to test whether age variable is significant, we can see the p-value corresponding to age(0.0359). Age variable is also significant since That is, the null hypothesis that the coefficient corresponding to age is 0 is reject with significance level 0.05.

We can also construct confidence interval of the probability of survival for each person with specific age of sex. Based on the lecture note, the following function computes the estimated probabilities and its confidence intervals

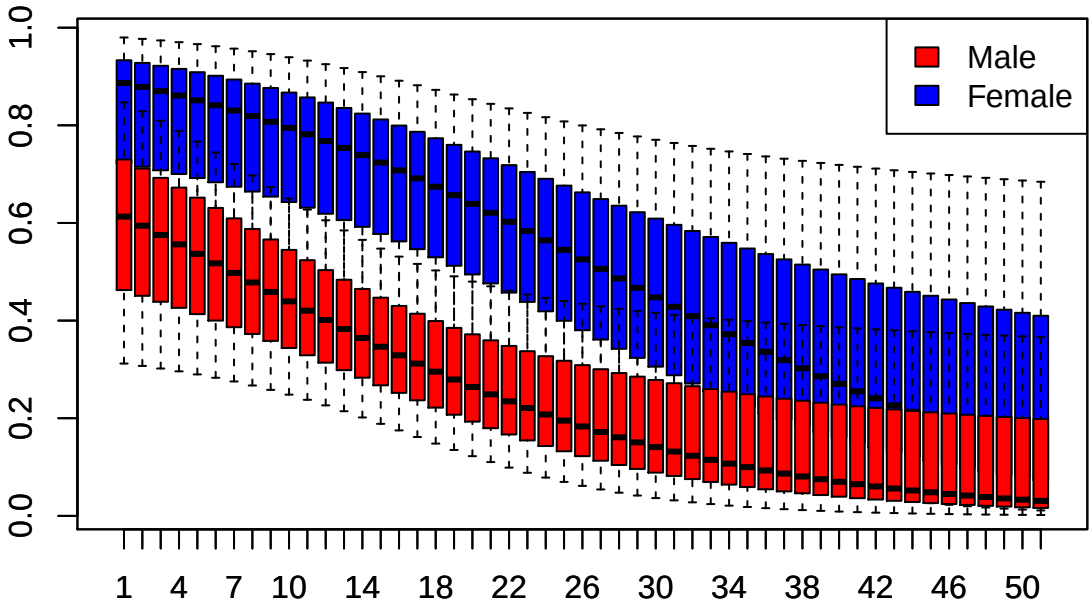
```r
conf2 <- function(x){
  b <- coef(model)
  p = 1/(1+exp(-t(x) %*% b))
  sexb = sqrt(t(x)%*% vcov(model) %*% x)
  cixb <- c(t(x) %*% b - 2 *sexb, t(x) %*% b + 2 *sexb)
  bound <- 1 / (1 + exp(-cixb))
  return(c(bound[1], p, bound[2]))
}
```

From this function, the the estimated probabilities and its confidence intervals are plotted

```r
age <- 15:65
res <- sapply(age, function(x) conf2(c(1, x, 0)))
boxplot(res, col = "blue")
```

```r
res2 <- sapply(age, function(x) conf2(c(1, x, 1)))
boxplot(res2, add = T, col = "Red")

legend("topright", c("Male", "Female"), fill = c("red", "blue"))
```



In this plot, each size of box is not meaningful but upper limit of box plot refers to the 95% upper bound for the confidence interval and lower limit of box plot refers to the the 95% lower bound for the confidence interval.

From the preceding analysis, we can say that males were more susceptible to death than females and the younger people were more able to survive than the older people.

Problem 4

Let y_A, y_B, y_C be the response for each experimental unit with treatment A, B, C respectively. The variance of each response is $Var(y_A) = 5, Var(y_B) = 5, Var(y_C) = 0.95$. However we can expect that the estimated variance $\hat{\sigma}^2$ would be underestimated because of large number of experimental units for treatment C. This may result in a larger value of t-statistics(or Z statistics) than what it should be, so it may reject the null hypothesis even when there is no significant difference between A and B.

To be specific, we can evaluate the expected value of mse under the suggested Gauss Markov linear model:$y \sim trt$. The expected value of mse is the weighted mean

$$\frac{9 * 5 + 9 * 5 + 49 * 0.95}{9 + 9 + 49} = 2.03806$$

However, when computing t-statistics for $trt_A - trt_B$, the expected value of the mse of the denominator is $Var(y_A) = Var(y_B) = 5$.

Hence, the mse is underestimated under the suggested model, so the null hypothesis might be reject even when the null hypothesis is true.

In short, the model has possible significant lack of fit so it might not be safe to go further on this model.