MDITERM: Statistical Computing (STT 802, EPI853b) 2022

10/26/2022

Your Name: _____

- This midterm is open book/notes but strictly individual.
- Please refer any questions that you may have to the instructor.

How to report your results and scripts

- I have uploaded in D2L the exam in three formats: Rmd, pdf, and word.
- If you choose to work with RStudio, you can provide your exam in one file (compiled to either pdf or html) that displays your script, results, and answers. You can use the Rmd as as a template, simply add your code, and answers and compile.
- If you decide to upload a word document, simply paste your code, outputs, and answers on the word document provided.

Submission

- You should submit your exam file in D2L by 4:20pm.
- Exams uploaded after 4:25 will recieve a 5 point penalty.
- The submission folders won't be avilable after 4:30pm.

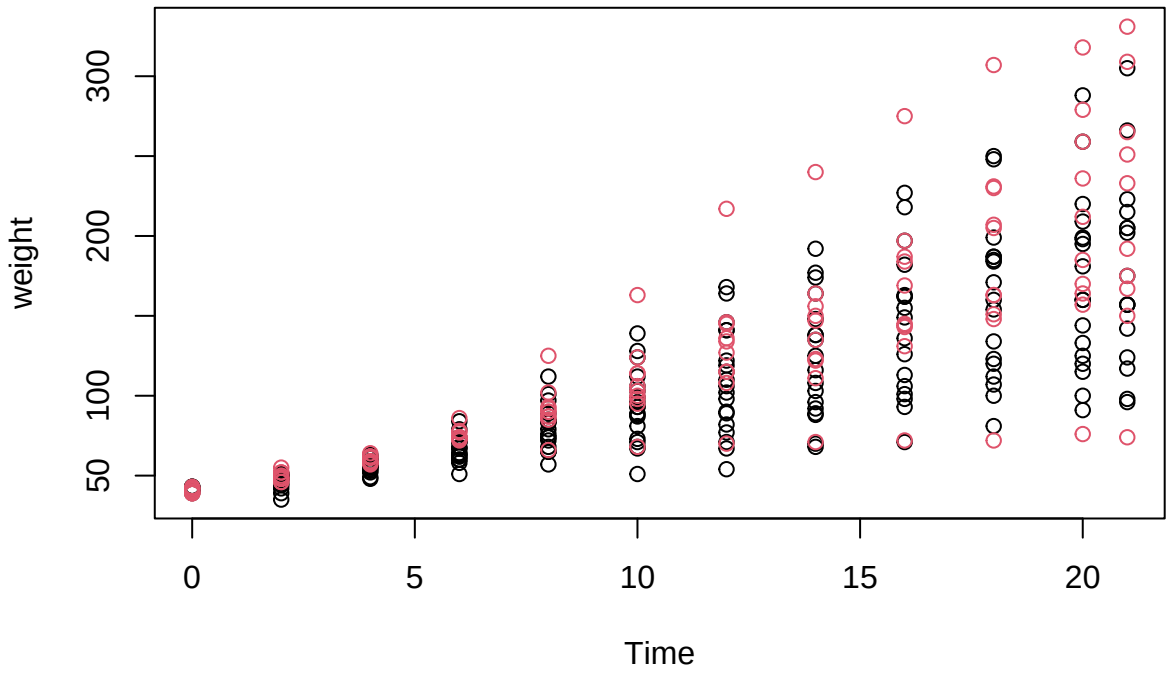
Questions

The exam has two questions (50 points each), each question has several sub-items.

1) Estimating growth curves using linear models and splines

The following data set has daily weight records of chickens that recieved two different diets.

```
rm(list=ls())
fname='https://raw.githubusercontent.com/vincentarelbundock/Rdatasets/master/csv/datasets/ChickWeight
chicks=read.csv(fname)
chicks=chicks[chicks$Diet%in%c(1,2),]
#xtabs(~chicks$Diet+chicks$Time)
plot(weight~Time,data=chicks,col=chicks$Diet)
```



1.1) Fit a linear model for weight as a function of time and diet. Consider the possiblity that diet may affect both the intercept and the slope of the regression.

- Does diet affect weight?
- Summarize your conclussions.

Solution:

- First note that diet has two levels and must be treated as a factor.
- According to the instructions, we should contemplate the possiblity that diet affects both the intercept and the slop (regression on time).
- Therefore, this will be a 2DF test with the folliwng null and alternative hypotheses

$H_0 : weight_{ij} = \mu + time_{ij}\beta + \varepsilon_{ij}$ $H_A : weight_{ij} = \mu + time_{ij}\beta_1 + D_{ij}\beta_2 + (D_{ij} \times time_{ij})\beta_3 + \varepsilon_{ij}$

Where D_{ij} is a dummy variable for one of the diets (by default in R, it will be for diet 2).

- Because this is a 2-df tests, we use an F-test. `anova()` can be used to perfomr such test.

```
H0=lm(weight~Time,data=chicks)
HA=lm(weight~Time*factor(Diet),data=chicks)
anova(H0,HA)
```

```
## Analysis of Variance Table
##
## Model 1: weight ~ Time
## Model 2: weight ~ Time * factor(Diet)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    338 469862
## 2    336 437076  2     32786 12.602 5.278e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- We reject the null.

Now, we inspect the results for H_A

```r
summary(HA)
```

```
##
## Call:
## lm(formula = weight ~ Time * factor(Diet), data = chicks)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -135.425  -16.780   -0.853   11.284  130.391
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         30.9310     4.4961   6.880 2.95e-11 ***
## Time                 6.8418     0.3608  18.963  < 2e-16 ***
## factor(Diet)2       -2.2974     7.6938  -0.299  0.76543
## Time:factor(Diet)2   1.7673     0.6052   2.920  0.00373 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.07 on 336 degrees of freedom
## Multiple R-squared:  0.6748, Adjusted R-squared:  0.6719
## F-statistic: 232.4 on 3 and 336 DF,  p-value: < 2.2e-16
```

- We see that there is a significant effect of diet on the slope. However, effect of diet on the intercept is not significant.

Conclusion: From the F-test (anova) we have strong evidence that diet affects growth. From the summary(HA) we see that diet does not affect the intercetp (if chickens were randomize to diets it makes sense that on average they would have similar weigths at the begining). However, there is a significant effect of diet in daily gain (slope) with Diet 1 giving lower daily gain (6.84 g/day) than diet 2 (8.61 g/day).

1.3) Non-linear growth? A researcher wants to learn wehather we have evidence of non-linear growth under Diet 1.

Using the data for Diet 1 only, test for non-linearity using a cubic B-spline with 4 DF.

```r
library(splines)
DATA=chicks[chicks$Diet==1,]
H0=lm(weight~Time,data=DATA)
HA=lm(weight~bs(Time,intercept=FALSE,df=4,degree=3),data=DATA)
anova(H0,HA)
```

```
## Analysis of Variance Table
##
## Model 1: weight ~ Time
## Model 2: weight ~ bs(Time, intercept = FALSE, df = 4, degree = 3)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    218 235212
## 2    215 229543  3    5668.4 1.7698 0.1539
```

Conclusion: The F-test shows a non-significant result. Therfore, we do not reject the linear growht model in favor of a non-liear one.

2) Maximum likelihood estimation

The density of a Beta distributed random variable has the following form

$f(y_i|\alpha_1, \alpha_2) = B(\alpha_1, \alpha_2)^{-1} \times y_i^{\alpha_1 - 1}(1 - y_i)^{\alpha_2 - 1}$

where $y \in [0, 1]$ is the random variable, $\alpha_1 > 0$ and $\alpha_2 > 0$ are shape parameters (denoted as shape1 and shape2 in `rbeta()`, `pbeta()`, `dbeta()`), and $B(\alpha_1, \alpha_2)$ is the Beta function (this funciton can be evaluated using the `beta()` function in R).

Therefore, the likleihood function for a random sample of n IID Beta distributed random variables is

$L(\alpha_1, \alpha_2|y_1, ..., y_n) = B(\alpha_1, \alpha_2)^{-n} \Pi_{i=1}^{n} y_i^{\alpha_1 - 1}(1 - y_i)^{\alpha_2 - 1}$

2.1) Write a function that evaluates the negative log-likelihood of random sample of IID Beta dstributed random variables.

Show your code here:

```
negLogLik=function(y,theta){
  -sum( dbeta(y,shape1=theta[1],shape2=theta[2],log=TRUE))
}

negLogLik2=function(y,theta){
  n=length(y)
  tmp= -n*log(beta(theta[1],theta[2]))+(theta[1]-1)*sum(log(y))+(theta[2]-1)*sum(log(1-y))
  return(-tmp)
}
```

2.2) Using the following data,

```
y=scan('~/Dropbox/STAT_COMP/2022/beta.txt',quiet=TRUE)
```

estimate the shape parameters via maximum likelihood.

Hint: use $\alpha_1 = 1$ and $\alpha_2 = 1$ as starting values.

- Did the algorithm converged?
- Report your parameter estimates.

```
fm=optim(fn=negLogLik,y=y,par=c(1,1))

fm$convergence # 0 indicates successful convergence
```

```
## [1] 0
```

```
fm$par # estimates
```

```
## [1] 10.676728  4.089861
```

```
fm2=optim(fn=negLogLik2,y=y,par=c(1,1))
fm2$convergence # 0 indicates successful convergence
```

```
## [1] 10
```

```r
fm2$par # estimates
```

```
## [1] 10.676729  4.089861
```

2.3) Provide SEs, and 95% CIs for each of the parameters.

```r
fm=optim(fn=negLogLik,y=y,par=c(1,1),hessian=TRUE)
V=solve(fm$hessian)
SE=sqrt(diag(V))
ANS=cbind('estimate'=fm$par,'SE'=SE,'LB'=fm$par-1.96*SE,'UB'=fm$par+1.96*SE)
round(ANS,2)
```

```
##      estimate   SE   LB    UB
## [1,]    10.68 1.51 7.71 13.64
## [2,]     4.09 0.56 3.00  5.18
```