STSB6806 Test 3 of 2021

Mathematical Statistics and Actuarial Science; University of the Free State

2021/07/29

## Time: 180 minutes; Marks: 50

# MEMORANDUM

# Instructions

* Answer all questions in a single R Markdown document. Please knit to PDF or Word at the end and submit both the PDF/Word document and the .Rmd file for assessment, in that order.
* Label questions clearly, as it is done on this question paper.
* All results accurate to about 3 decimal places.
* Show all derivations, formulas, code, sources and reasoning.
* Intervals should cover 95% probability unless stated otherwise.
* No communication software, no devices, and no communication capable websites may be accessed prior to submission. You may not (nor even appear to) attempt to communicate or pass information to another student.

# Introduction

The data is provided on <https://ufs.blackboard.com>. It is about how much yield a farmer can get from a field of fixed size. Two types of fertiliser are being tested, a Treatment fertiliser and a Control fertiliser. Each fertiliser is applied to a number of different fields and with three levels of watering (Low, Medium, High).

# Question 1

**1.1)** Read in the data set and explore it visually by drawing a box plot of Yield against the interaction of Treatment and Watering. **[ 6 ]**

openxlsx::read.xlsx("Test3Data.xlsx", "Test3Data") -> d  
par(mar=c(5,5,1,1))  
boxplot(Yield~Treatment+Watering, data=d, cex.axis=0.5, col=c('red','blue'))





###### Load data [2], Box plot [2], Discussion [2]

**1.2)**  
Perform a simple three way ANOVA of Yield on Treatment, Field, and Watering. Which effects are statistically significant? **[ 3 ]**

kable(anova(lm1 <- lm(d$Yield ~ Treatment + Field + Watering, data=d)))

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
| Treatment | 1 | 376.4544 | 376.45441 | 26.764571 | 0.0000082 |
| Field | 7 | 253.5745 | 36.22493 | 2.575464 | 0.0287507 |
| Watering | 2 | 345.0923 | 172.54617 | 12.267420 | 0.0000818 |
| Residuals | 37 | 520.4198 | 14.06540 | NA | NA |

###### ANOVA [2], Saying all significant [1]

**1.3)**  
What is the estimated treatment effect? As in, how much higher would you expect the yield to be according to the linear model including the three effects as fixed effects? Also give a 95% interval for the estimated treatment effect. **[ 4 ]**

kable((lm1coef <- coef(summary(lm1 <- lm(d$Yield ~ Treatment + Field + Watering, data=d)))[2,]))

|  |  |
| --- | --- |
|  | x |
| Estimate | 5.6010000 |
| Std. Error | 1.0826434 |
| t value | 5.1734487 |
| Pr(>|t|) | 0.0000082 |

c(lower=lm1coef[1] - lm1coef[2]\*qt(0.975,37), upper=lm1coef[1] + lm1coef[2]\*qt(0.975,37))

| lower.Estimate upper.Estimate   
| 3.407356 7.794644

###### About 5.6 higher [2], Interval about 3.4 to 7.8 [2]

**1.4)**  
Test (visually or using a standard hypothesis test) whether the residuals have constant variance across the three levels of Watering. Is the variance of the ‘Low’ observations the same as that of the ‘Medium’ and ‘High’ groups? **[ 3 ]**

boxplot(resid(lm1)~d$Watering, xlab = 'Watering level', ylab = 'ANOVA residuals', col=rainbow(3))



bartlett.test(resid(lm1)~d$Watering)

|   
| Bartlett test of homogeneity of variances  
|   
| data: resid(lm1) by d$Watering  
| Bartlett's K-squared = 10.373, df = 2, p-value = 0.005593

###### Box plot or Bartlett/Levene test [2], saying that the variances differ [1].

**1.5)**  
Redo the linear model, but allowing each group (Low/Medium/High) to have a different variance in the model. Make the Field effect a random effect in the model. Estimate the treatment effect again, and 95% interval. Test whether it has changed. **[ 17 ]**

library(rstan)  
mycores <- 3  
options(mc.cores = mycores)

// This Stan block defines a 3 factor model with changing variation between groups, by Sean van der Merwe, UFS  
data {  
 int<lower=1> n; // number of observations  
 int<lower=1> ngrp; // number of groups  
 int<lower=1> neffect; // number of effects  
 real y[n]; // observations  
 int<lower=1,upper=ngrp> grp[n]; // group membership  
 int<lower=0,upper=1> treat[n]; // treatment received (1) versus control (0)  
 int<lower=1,upper=neffect> effect[n]; // effect membership  
}  
parameters {  
 real betaTreat; // Treatment effect   
 real betaGroup[ngrp]; // Group effect  
 real sGroup[ngrp]; // Group standard effects on log scale  
 real betaEffect[neffect]; // Random effect  
 real<lower=0> tau; // Random effect standard deviation  
}  
transformed parameters {  
 real mu[n]; // expected values  
 real<lower=0> s[n]; // standard deviations  
 for (i in 1:n) {  
 mu[i] = betaTreat\*treat[i] + betaGroup[grp[i]] + betaEffect[effect[i]];  
 s[i] = exp(sGroup[grp[i]]);  
 }  
}  
model {  
 y ~ normal(mu, s); // fit the data pattern  
 betaEffect ~ normal(0, tau);  
}

saveRDS(LM1, file = 'LM1.Rds')

stan\_data1 <- list(n=length(d$Yield), y=d$Yield, ngrp=3, grp=as.numeric(factor(d$Watering)), effect=as.numeric(factor(d$Field)), neffect=8, treat=((d$Treatment=="Treatment") + 0))  
ModelFit1 <- sampling(LM1, stan\_data1, iter = 20000, chains = mycores)

saveRDS(ModelFit1, file = 'LM1sim.Rds')

draws1 <- extract(ModelFit1)  
kable(summary1 <- round(summary(ModelFit1)$summary[1:16,c(-5,-7)],3))

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | mean | se\_mean | sd | 2.5% | 50% | 97.5% | n\_eff | Rhat |
| betaTreat | 5.430 | 0.005 | 0.809 | 3.825 | 5.425 | 7.029 | 23277.627 | 1.000 |
| betaGroup[1] | 13.613 | 0.014 | 1.665 | 10.307 | 13.579 | 16.934 | 14020.242 | 1.000 |
| betaGroup[2] | 7.131 | 0.011 | 1.027 | 5.102 | 7.109 | 9.167 | 9004.421 | 1.001 |
| betaGroup[3] | 11.331 | 0.012 | 1.287 | 8.773 | 11.320 | 13.900 | 10971.945 | 1.000 |
| sGroup[1] | 1.684 | 0.001 | 0.199 | 1.316 | 1.674 | 2.106 | 19131.386 | 1.000 |
| sGroup[2] | 0.651 | 0.002 | 0.222 | 0.249 | 0.641 | 1.115 | 9745.493 | 1.001 |
| sGroup[3] | 1.252 | 0.001 | 0.215 | 0.845 | 1.249 | 1.694 | 21961.203 | 1.000 |
| betaEffect[1] | 1.390 | 0.018 | 1.339 | -0.942 | 1.290 | 4.294 | 5449.543 | 1.002 |
| betaEffect[2] | -2.732 | 0.031 | 1.456 | -5.733 | -2.698 | -0.153 | 2159.518 | 1.001 |
| betaEffect[3] | -0.423 | 0.011 | 1.211 | -2.985 | -0.361 | 1.855 | 11462.588 | 1.000 |
| betaEffect[4] | -0.461 | 0.011 | 1.196 | -2.952 | -0.407 | 1.866 | 12450.149 | 1.000 |
| betaEffect[5] | 0.063 | 0.011 | 1.177 | -2.289 | 0.042 | 2.458 | 12275.515 | 1.000 |
| betaEffect[6] | 1.236 | 0.018 | 1.246 | -1.024 | 1.192 | 3.844 | 5022.326 | 1.002 |
| betaEffect[7] | 0.975 | 0.013 | 1.231 | -1.279 | 0.893 | 3.622 | 9503.489 | 1.001 |
| betaEffect[8] | -0.078 | 0.011 | 1.211 | -2.522 | -0.055 | 2.368 | 12714.809 | 1.001 |
| tau | 1.991 | 0.023 | 1.012 | 0.371 | 1.829 | 4.435 | 1978.950 | 1.001 |

pvalfunc <- function(sims,target=0) { 2\*min(mean(sims<target),mean(sims>target)) }  
cat('\n\n\*From the simulation summary,\* \*\*the estimated treatment effect is ', summary1[1,1], ' with interval (', summary1[1, 4], '; ', summary1[1, 6], ') and the p-value equivalent of comparing it to the previous estimate is ', round(pvalfunc(draws1$betaTreat - lm1coef[1]),3), ', which is not significant.\*\*\n\n', sep='')

*From the simulation summary,* **the estimated treatment effect is 5.43 with interval (3.825; 7.029) and the p-value equivalent of comparing it to the previous estimate is 0.823, which is not significant.**

###### Field as random effect [4], grouped variances [5], Treatment effect incorporated correctly without any identifiability issues [2], model well fit [2], treatment effect and interval given [2] and showing that it has not changed [2].

## Prior information

Expert information from other studies suggest that the yield under low water conditions should be about 6 lower than with high watering, and under medium watering it should be about 2 lower than with high watering. All other effects being zero, high watering should result in an average yield of about 14, give or take 0.5 with 70% certainty.

It is also suggested that the standard deviation between fields should vary with an average of 1.5 and a variance of 2.

**1.6)** Incorporate this prior knowledge into your model and discuss the impact on all modelled fixed effects, particularly the treatment effect. Note that you may incorporate the prior information in any form that agrees with what was given, not necessarily exactly as supplied.**[ 9 ]**

// This Stan block defines a 3 factor model with changing variation between groups, by Sean van der Merwe, UFS  
data {  
 int<lower=1> n; // number of observations  
 int<lower=1> ngrp; // number of groups  
 int<lower=1> neffect; // number of effects  
 real y[n]; // observations  
 int<lower=1,upper=ngrp> grp[n]; // group membership  
 int<lower=0,upper=1> treat[n]; // treatment received (1) versus control (0)  
 int<lower=1,upper=neffect> effect[n]; // effect membership  
 real priormu[ngrp]; // prior expected values  
 real priorsigma[ngrp]; // prior standard deviations  
 real prioralpha; // prior on tau  
 real priorlambda; // prior on tau  
}  
parameters {  
 real betaTreat; // Treatment effect   
 real betaGroup[ngrp]; // Group effect  
 real sGroup[ngrp]; // Group standard effects on log scale  
 real betaEffect[neffect]; // Random effect  
 real<lower=0> tau; // Random effect standard deviation  
}  
transformed parameters {  
 real mu[n]; // expected values  
 real<lower=0> s[n]; // standard deviations  
 for (i in 1:n) {  
 mu[i] = betaTreat\*treat[i] + betaGroup[grp[i]] + betaEffect[effect[i]];  
 s[i] = exp(sGroup[grp[i]]);  
 }  
}  
model {  
 y ~ normal(mu, s); // fit the data pattern  
 betaEffect ~ normal(0, tau);  
 betaGroup ~ normal(priormu, priorsigma);  
 tau ~ gamma(prioralpha, priorlambda);  
}

saveRDS(LM2, file = 'LM2.Rds')

stan\_data2 <- list(n=length(d$Yield), y=d$Yield, ngrp=3, grp=as.numeric(factor(d$Watering)), effect=as.numeric(factor(d$Field)), neffect=8, treat=((d$Treatment=="Treatment") + 0), priormu=c(14, 8, 12), priorsigma=rep(0.5,3), prioralpha=((1.5)^2)/2, priorlambda=1.5/2)  
ModelFit2 <- sampling(LM2, stan\_data2, iter = 20000, chains = mycores)

saveRDS(ModelFit2, file = 'LM2sim.Rds')

draws2 <- extract(ModelFit2)  
kable(round(summary(ModelFit2)$summary[1:4,c(-5,-7)],3))

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | mean | se\_mean | sd | 2.5% | 50% | 97.5% | n\_eff | Rhat |
| betaTreat | 5.109 | 0.005 | 0.714 | 3.695 | 5.112 | 6.508 | 21650.64 | 1 |
| betaGroup[1] | 14.009 | 0.003 | 0.476 | 13.085 | 14.010 | 14.942 | 27887.37 | 1 |
| betaGroup[2] | 7.783 | 0.003 | 0.409 | 6.977 | 7.782 | 8.589 | 23366.23 | 1 |
| betaGroup[3] | 11.949 | 0.003 | 0.448 | 11.074 | 11.946 | 12.827 | 28802.42 | 1 |

###### Model code correctly adapted to include a prior on each group effect [3], and a prior on the standard deviation between fields [2], then results showing that the group effects are now closer to the prior values [2], and a discussion explaining that the estimated treatment effect is reduced [2], making the model more conservative.

**1.7)** Create predictions and intervals for the yield from a random future field under medium watering, one set for the treatment and one set for the control.**[ 6 ]**

nsims <- length(draws2$betaTreat)  
newfieldeffects <- rnorm(nsims, 0, draws2$tau)  
newmus <- newfieldeffects + draws2$betaGroup[,3]  
newsigmas <- exp(draws2$sGroup[,2])  
controlsims <- rnorm(nsims, newmus, newsigmas)  
treatmentsims <- rnorm(nsims, newmus + draws2$betaTreat, newsigmas)  
results <- rbind(t(quantile(treatmentsims, c(0.025, 0.5, 0.975))), t(quantile(controlsims, c(0.025, 0.5, 0.975))))  
colnames(results) <- c('Lower', 'Median', 'Upper')  
kable(data.frame(Treatment=c('Treatment', 'Control'), round(results,3), row.names = NULL))

|  |  |  |  |
| --- | --- | --- | --- |
| Treatment | Lower | Median | Upper |
| Treatment | 11.617 | 17.064 | 22.577 |
| Control | 6.717 | 11.942 | 17.212 |

###### Generating random future fields using all relevant parameters [2], generating random yields from each generated field using all relevant parameters [2], constructing predictions and intervals [2].

**1.8)** How likely is it to get lower yield under the treatment than under the control, in these circumstances?**[ 2 ]**

cat('The probability of getting lower yield under treatment is about ', round(mean(treatmentsims < controlsims),3)\*100, '% in a random future field under medium watering.', sep='')

| The probability of getting lower yield under treatment is about 4% in a random future field under medium watering.

## Points total

The points on the test add up to **50**