Regression Methods: Problems MATH-408
Anthony Davison

Problem 1 Figure 1 shows hourly rainfall amounts (mm) at a weather station in Eskdalemuir, Scotland. The code below (and in the file `esk.R`) imports these data and performs an analysis to assess seasonality in the daily rainfall probabilities and amounts. Use that code to perform the suggested analysis, and then do the further analyses suggested at the end.

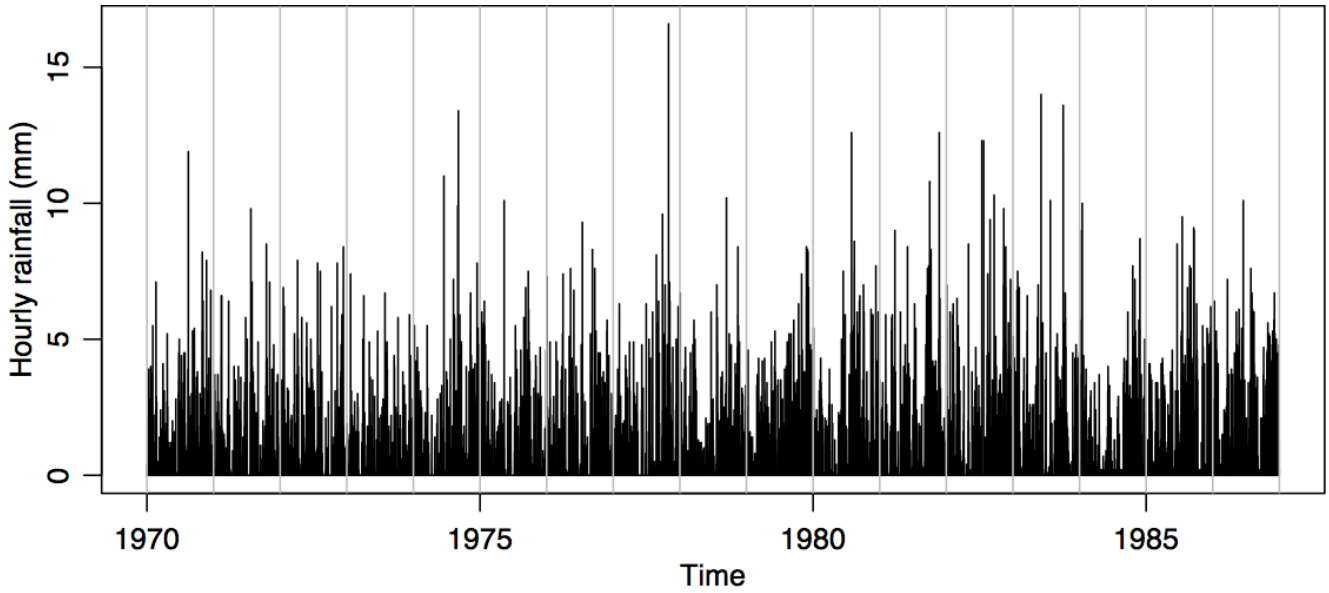

Figure 1: Hourly rainfall (mm) in Eskdalemuir

.

```
# read and prepare data

load(file="Eskdalemuir-data.Rdata")

# drop 29 February for leap years and create time series with 365*24 observations per period

rain.matrix <- subset(h.precip.matrix, subset=!((h.precip.matrix[,2]==2)&(h.precip.matrix[,3]==29)))

esk.rain <- ts(rain.matrix[,4],start=1970,frequency=365*24)

# plot data

plot(esk.rain,ylab="Hourly rainfall (mm)")
abline(v=1970+c(0:17),col="grey")

# create hourly and daily variables, and check on their plausibility

hour <- rep(0:23,6205)
day <- rep(1:365,each=24,length.out=length(esk.rain))

plot(0:23,tapply(esk.rain,hour,mean))
plot(1:365,tapply(esk.rain,day,mean))

# take daily totals, owing to strange variation around 8-9 am

temp <- matrix(esk.rain,nrow=365*17,ncol=24,byrow=T)
esk <- ts(apply(temp,1,sum),start=1970,frequency=365)
day <- rep(1:365,length.out=length(esk))

# Binary model for rainfall occurrence ====================================

# first model has different rainfall probabilities each day, not very sensible

fit1 <- glm((esk>0)~factor(day),family=binomial)
anova(fit1)
summary(fit1)
esk.fit1 <- fitted(fit1)

# normal scores plot of residuals.  Discuss!

esk.res1 <- residuals(fit1)
qqnorm(esk.res1)
```

```
#fit harmonic terms for time of year

fit2 <- glm((esk>0)~sin(2*pi*day/365)+cos(2*pi*day/365)+sin(2*pi*2*day/365)+cos(2*pi*2*day/365),family=binomial)
anova(fit2)
summary(fit2)
esk.fit2 <- fitted(fit2)

# Note that the sin and cos terms should enter in the ANOVA in pairs for each harmonic, since it makes no sense to consider them individually

# would one set of harmonic terms be enough?  Do we need more than two sets?

# Is the fit OK?  Is there overdispersion?  If so, try a quasi=likelihood fit

# Gamma model for rainfall amounts ======================================

# look at data

hist(esk[esk>0],prob=T,nclass=100)

# fit simple model, use F test for analysis of deviance

fit5 <- glm(esk~factor(day),family=Gamma(link="log"),subset=(esk>0))
anova(fit5,test="F")
summary(fit5)
esk.fit5 <- fitted(fit5)

# Fitted values for this fit.  What do you think?

plot(day[esk>0],esk[esk>0],xlab="Day of year",ylab="Rainfall (mm)",xlim=c(1,365),pch=16,cex=0.5)
points(1:365,esk.fit5[1:365],col="red",pch=16)

# normal scores plot of residuals.  What do you think?

esk.res5 <- residuals(fit5)
qqnorm(esk.res5)

# harmonic terms

fit6 <- glm(esk~sin(2*pi*day/365)+cos(2*pi*day/365),family=Gamma(link="log"),subset=(esk>0))

anova(fit6)
summary(fit6)
esk.fit6 <- fitted(fit6)

plot(day[esk>0],esk[esk>0],xlab="Day of year",ylab="Rainfall (mm)",xlim=c(1,365),pch=16,cex=0.5)
points(1:365,esk.fit5[1:365],col="red",pch=16)
lines(1:365,esk.fit6[1:365],col="blue")

# generalized additive models (GAMs) for the adventurous (or for later, once we've done them in class ....) ============================

# Now we fit a GAM to the occurence data, with a cyclic spline for day of the year

library(mgcv)
fit3 <- gam((esk>0)~s(day,bs="cc"),family=binomial)
anova(fit3)
summary(fit3)
esk.fit3 <- fitted(fit3)

# Is this an equally good fit as the harmonic one, or better, or worse?
# add line to graph for days

lines(1:365,esk.fit3[1:365],col="blue")

# now check whether a GAM gives better results than the harmonic model for the sizes

# for both analyses above, look and see if allowing for changes over the years improves the fit by a lot

# Hint: either create a variable year and then a variable for the total days since the start of the data,
# or (maybe less good) chekc out use of the 'by' option in a smooth used in gam()
```