Stat 217 Project 3

Due: Friday, October 31st, beginning of class

(Old Faithful): Old Faithful Geyser in Yellowstone National Park derives its names and fame from the regularity (and beauty) of its eruptions. Rangers usually post the predicted times of eruptions for visitors. R. A. Hutchinson, a park geologist, collected measurements of the eruption durations (in minutes) and the subsequent time intervals before the next eruption (in minutes) over an 8-day period. Help rangers use the data to explain the relationship between duration and subsequent time to the next eruption. Then, help them use that relationship to predict when next eruptions will occur. Provide the ranger with an estimate of the mean length of time until the next eruption after one lasting for 2 minutes. Also, provide the ranger with a prediction of how many minutes visitors will have to wait after a future eruption lasting 4 minutes. Be sure to give the rangers appropriate uncertainty intervals to go with estimates and/or predictions. Make sure to state whether your intervals are confidence intervals or prediction intervals.

Write a statistical report, following the project writing guidelines. For your *Summary of Statistical Findings*, include 4 sentences:

1. An evidence sentence for the relationship between eruption duration and waiting time
2. An "it is estimated" sentence to describe this relationship (include uncertainty interval)
3. One sentence to describe the mean length of time until the next eruption after one lasting for 2 minutes (include uncertainty interval)
4. One sentence to describe your prediction of how many minutes visitors will have to wait after a future eruption lasting 4 minutes (included uncertainty interval)

You will have to download the Old Faithful Data from D2L and use the data import wizard to load it into Rstudio. Use the following R-code:

```
with(faith.data,plot(INTERVAL~DURATION))

lm.out <- lm(INTERVAL~DURATION, data=faith.data)
summary(lm.out)

with(faith.data, plot(DURATION,INTERVAL,type="n", xlim=c(1,6),
                      main="Waiting time vs. Duration"))
with(faith.data, points(DURATION,INTERVAL, pch=16))
abline(lm.out)
abline(h=seq(40,90,by=10), lty=2)

par(mfrow=c(2,2))
plot(lm.out)

confint(lm.out)
dur.2 <- with(faith.data, data.frame(DURATION=2))
fit.2 <- predict(lm.out, newdata=dur.2, se.fit=TRUE, interval="confidence",level=0.95)
```

After your Scope of Inference, describe what you see in the following plot in one additional paragraph.

```
### The following is the code for the plot with 95% confidence interval and
###prediction interval bands
###Highlight this whole chunk of code and run it all at once

new <- data.frame(faith.data$DURATION = seq(1.7, 4.9, length=107))
#50 values between -220 and 1090
est.mean.cis <- predict(lm.out, newdata=new, interval="confidence")
pred.pis <- predict(lm.out, newdata=new, interval="prediction")

## Make a confidence BAND using the Scheffe multiplier
est.mean.out <- predict(lm.out, newdata=new, se.fit=TRUE, interval="confidence")
est.mean.ses <- est.mean.out$se.fit
look <- est.mean.out$fit
est.means <- est.mean.out$fit[,1] #takes first column of the $fit matrix
#which are the fitted means for each of the 50 values
sch <- (sqrt(2*qf(.95,2,105))) #this is the Scheffe multiplier
tmult <- qt(.975,105)  #compare to the t-multiplier
conf.BAND.Scheffe.low <- est.means - (sqrt(2*qf(.95,2,105)))*est.mean.ses
conf.BAND.Scheffe.hi <- est.means + (sqrt(2*qf(.95,2,105)))*est.mean.ses
sch.band <- cbind(conf.BAND.Scheffe.low, conf.BAND.Scheffe.hi)


with(faith.data, plot(DURATION, INTERVAL, type="n", ylab="Interval (minutes)",
xlab="Duration (minutes)", main="Old Faithful Data"))

abline(lm.out, lwd=2) #fitted line
lines(new$DURATION, est.mean.cis[,1], lty=1, lwd=2)  #another way to add fitted line
lines(new$DURATION, est.mean.cis[,2], lty=2, lwd=2, col=2) #lower pointwise CI
lines(new$DURATION, est.mean.cis[,3], lty=2, lwd=2, col=2) #upper pointwise CI
lines(new$DURATION, pred.pis[,2], lty=1, lwd=2, col=3)  #lower prediction PI
lines(new$DURATION, pred.pis[,3], lty=1, lwd=2, col=3) #upper prediction PI
lines(new$DURATION, conf.BAND.Scheffe.low, lty=1, lwd=2, col=4)  #lower confidence band
lines(new$DURATION, conf.BAND.Scheffe.hi, lty=1, lwd=2, col=4)  #upper confidence band
with(faith.data, points(DURATION, INTERVAL, pch=16, cex=0.5))
```