HW 3: Sting data analysis

Paul Harmon and Steve Durtka

September 15, 2017

You must work in groups (2 to 4 students) or get an approved exemption.

Read the paper "Honey bee sting pain index by body location" by Michael Smith, published in the *PeerJ* in 2014 - posted on D2L along with a supplement they published.

A slightly cleaned up version of the data set that was extracted from the supplemental materials is available on D2L. Note that coded the first forearm measurement each day as *forearm1* and the second one I left as *forearm*. It contains contains all the observations including their calibration observations on the forearm each day.

```r
library(readr)
getwd()
```

```
## [1] "C:/Users/paulh/Documents/Mixed-Models-Bees"
```

```r
#reads in data as tibbles
sd_fixed <- read_csv("data/stingdata_fixed.csv")
sd_fixed$BLs <- sd_fixed$Body_Location
sd_fixed$BLs <- with(sd_fixed, reorder(Body_Location, Rating, median))
```

Now we will try to do a more complete analysis of the Body Location differences using these data.

1) **Describe the study design. Note the numbers of observations per round and day and number of days and any randomization that occured.**

-The author stung himself in 25 randomly-selected locations. Over the course of 38 days, the author stung himself in the forearm first, then stung three randomly-selected body locations, and finally stung himself on the forearm again for a total of five stings per day. During this period, three complete rounds of the 25 body locations were covered. These data constitute repeated measures on a single subject (the author) over time.

The randomly-selected locations were chosen using a program in R.

2) **With the design of this study, we could attempt to account for differences from day to day with a random effect. Why would we not expect too much day to day variation in the responses once we account for location in this study?**

We could involve a random effect to control for the day-to-day variation in pain ratings if all of the sting locations had measurements for each day. However, because only 5 randomly selected locations were used, only a few days had stings at the same sites for comparison.

Moreover, the author had stung himself repeatedly for over a month prior to the study, so it is reasonable that his body sytems had at least somewhat acclamized to repeated bee stings.

3) **Fit a linear model that just contains the location of the sting as an explanatory variable (do not account for day) as coded in the *BLs* variable. Perform an overall test for some difference in mean pain ratings across the locations and report the results in a sentence, carefully including all your evidence in the sentence - not in the R output. You will not match their results because they had a data coding issue in their analysis.**

The results of the model indicate evidence against the null hypothesis that mean pain ratings are the same for all locations; rather, an ANOVA F-test (with 25 and 94 df, respectively) suggests that bee stings in at least one different body location, on average, tend to be associated with different pain ratings than others.

```r
lm.1 <- lm(Rating~BLs, data = sd_fixed)
#summary(lm.1)
library(pander)
pander(anova(lm.1))
```

Table 1: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
BLs	25	248	9.919	27.42	7.938e-33
Residuals	94	34	0.3617	NA	NA

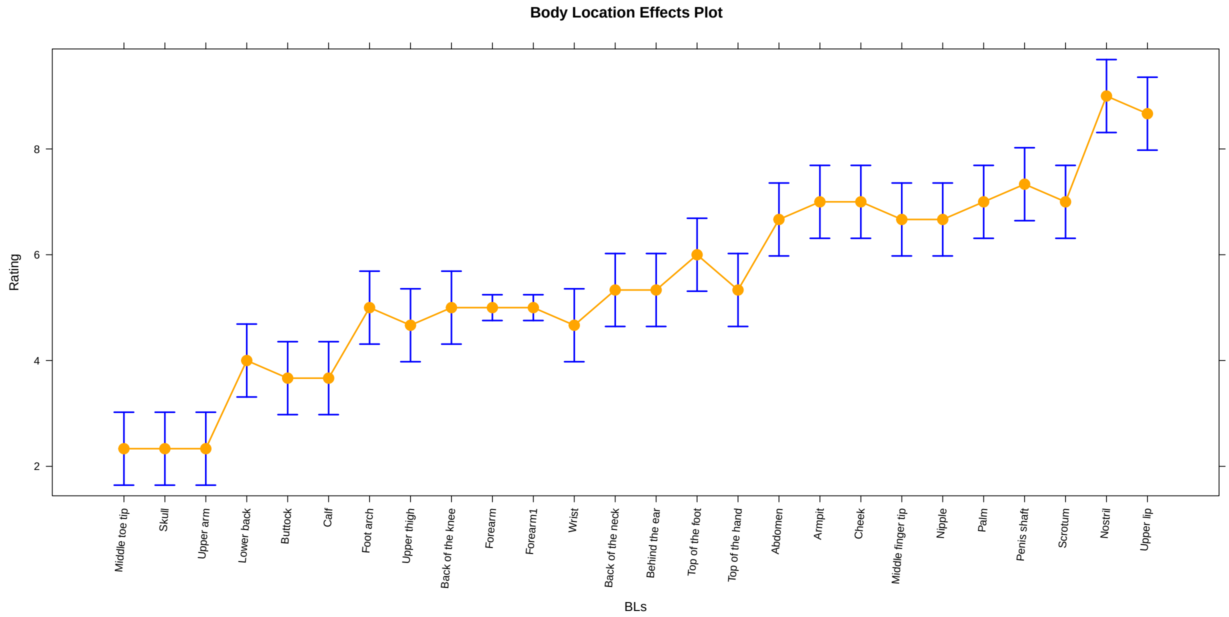
4) **Discuss scope of inference for this test. In other words, address whether you can make causal inferences/not and how far the inferences can extend - and make your discussion specific to the problem at hand.**

The author does a nice job of noting that the scope of inference of this study is limited just to him - inferences on a non-random sample of 1 do not allow for generalization to a wider population. That being said, the order of body part stings was random, indicating that we can at least conclude that the bee stings caused the author to feel different levels of pain at those different selected locations.

5) **Use the effects package to make a nice plot of the results from the model, using the `rotx` option to make the levels more readable such as with `plot(allEffects(modelname),rotx=90)`. Make sure you are using the `BLs` version of the explanatory variable when you fit the model above.**

The effects plot shows the estimated mean pain ratings for each body part (including the baseline body part, the middle toe tip). The plot is discussed in question 6.

```r
library(effects)
plot(allEffects(lm.1), rotx = 85, main = "Body Location Effects Plot",
     color = c("orange","blue"))
```



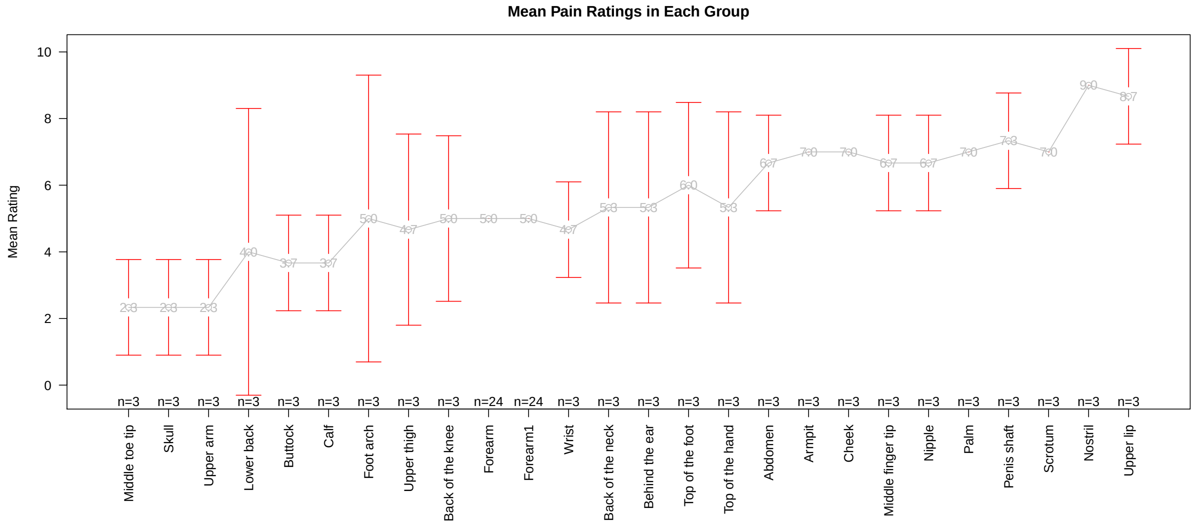
6) **Use the effects plot to discuss the mean pain ratings. This can be similar to how they discussed things in the paper.**

The effects plot shows the estimated mean pain ratings with standard errors plotted around each point estimate. The most painful locations on the body were the nostril and upper lip, as well as the genitalia-related areas of the body. The least-painful parts of the body were the middle toe tip, skull, and upper arm, with estimated mean pain ratings between 1 and 3. Pain ratings seem to be fairly consistent across the body as the standard errors for each location are not all that large (less than 1.0 on the pain rating scale on either side of the point estimate).

7) **We can also use the `plotmeans` function from `gplots` to explore data sets in one-way ANOVA situations. Compare these results to the results from the linear model.**

These ratings allow for different levels of variation for each body part, as opposed to the linear model that imposes constant standard errors for each group. We can see that the lower back, foot arch,upper thigh, back of the neck and behind the ear have the most variation in mean pain ratings. The body parts with the smallest and largest average pain ratings appeared to have fairly consistent (narrow) mean pain ratings relative to the other ones mentioned; these are the same as in the previous plot.

```
library(gplots)
par(mfrow = c(1,1))
par(mar = c(8,5,3,3))
plotmeans(Rating~BLs,data=sd_fixed,mean.labels=TRUE, digits=1,barcol="red",
          col="grey",xlab = '',ylab = 'Mean Rating', las = 2)
#axis(1,labels = unique(sd_fixed$BLs),at = 1:26,las = 2,cex = .5)
#rather than doing this I added 'las =2' to the above line of code,
#this gives the sorted groups without making us do it by hand
title("Mean Pain Ratings in Each Group")
```

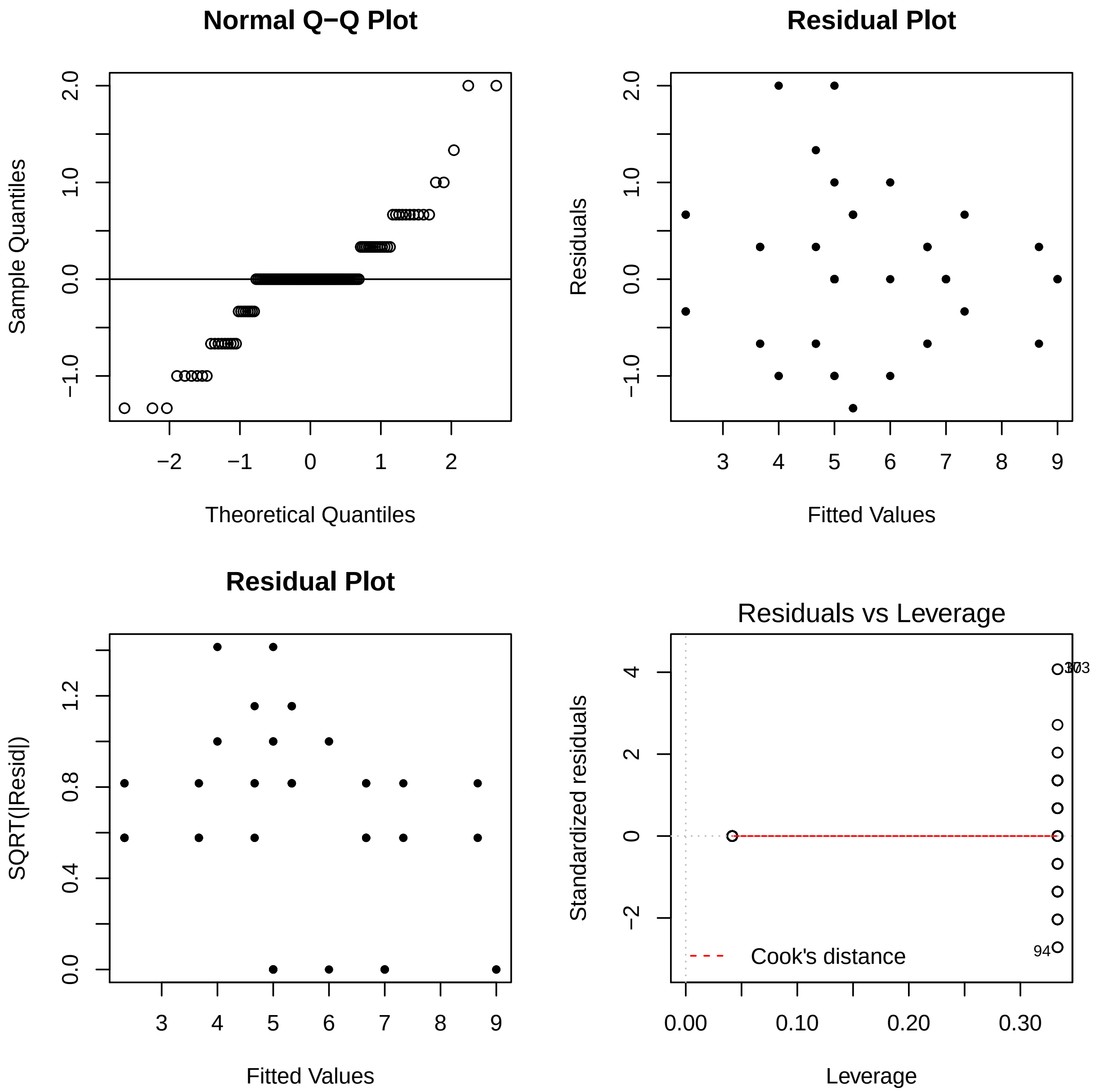


8) **Make our regular panel of four diagnostic plots and discuss the assumptions based on these plots.**

The QQ plot shows that the points fall roughly on the X=Y line; however, since the response is measured on a likert-type scale, it is discrete. This model treats the response as normal (and treats the discretized response as continuous), which does not seem unreasonable given this plot. However, a multinomial response might be a better option to consider.

The assumption of constant variance looks like it may be somewhat violated as the fitted values near the middle appear to have wider variation than those near the minimum and maximum fitted values. None of the observations appear to be influential points. This is a saturated model since the explanatory variables are the groups (locations) themselves, so we do not need to assess linearity in this model.

```r
par(mfrow=c(2,2))
par(mar = c(4,4,4,2))
qqnorm(resid(lm.1));qqline(lm.1$resid) #QQPlot
plot(fitted(lm.1),resid(lm.1), pch = 20,
     main = "Residual Plot",xlab = "Fitted Values",ylab = "Residuals")
plot(fitted(lm.1),sqrt(abs(resid(lm.1))), pch = 20,
     main = "Residual Plot",xlab = "Fitted Values",ylab = "SQRT(|Resid|)")
plot(lm.1, which = 5)
```



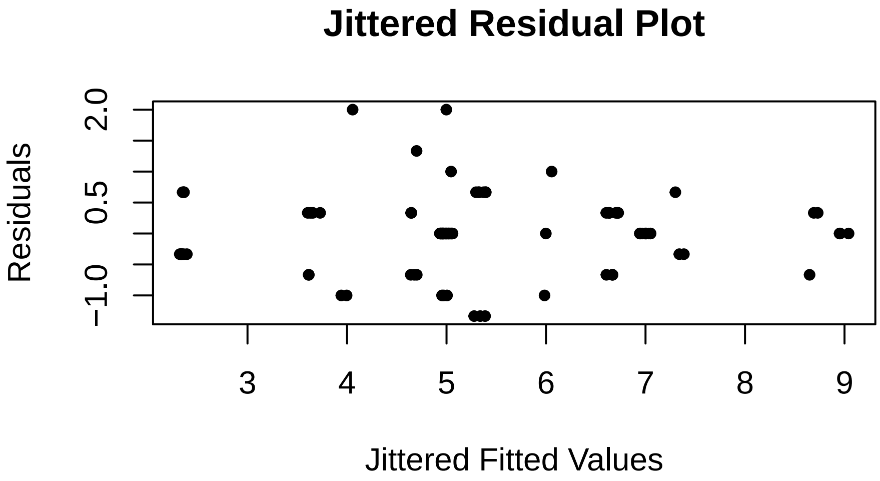
9) **Our regular diagnostic plots are a little misleading here. To make them better, we can "jitter" various aspects of the diagnostics to help us see if points are overplotting in the regular version of things. Use the following code to make a plot of the regular residuals vs jittered fitted values. Discuss the assumptions about the model that can be assessed using just this plot.**

- **Use something like `plot((residuals(MODELNAME))~jitter(fitted(MODELNAME)))`**

The jittered residual plot can be used to assess the non-constant variance of the model; however, it likely should not be used to assess linearity because it jitters observations that are observed at the same x-value over a range of x-values. Thus, it will tend to make non-linear relationships look different than they really are by flattening them out. (It often will not make a great deal of sense to evaluate linearity anyway since these models are often saturated as these are used for observations of group means (i.e. ANOVA), not quantatitive variables.

In any case, jittering the values makes this issue of non-constant variance look more severe as observations between fitted values 4-6 have much wider variation than the observations at larger fitted values. This makes sense given that the mean plots showed several body parts had wider variation in mean pain ratings than those with the smallest and largest pain ratings.

```r
plot(resid(lm.1)~jitter(fitted(lm.1)),pch = 20,
     main = "Jittered Residual Plot", xlab = "Jittered Fitted Values",
     ylab = "Residuals")
```


Jittered Residual Plot
Residuals
Jittered Fitted Values