HW6: Sections 7.1, 7.2, 7.3, 7.4.1, and 7.4.2

Solutions

The code below just loads some packages and makes it so that enough digits are printed that you won't get confused by rounding errors.

```
library(dplyr) # functions like summarize
library(ggplot2) # for making plots
library(mosaic) # convenient interface to t.test function
library(readr)
options("pillar.sigfig" = 10) # print 10 significant digits in summarize output
```

Brain activity in string instrument players: adapted from Sleuth3 Problem 7.28

Studies over the past two decades have shown that activity can effect the reorganization of the human central nervous system. For example, it is known that the part of the brain associated with activity of a finger or limb is taken over for other purposes in individuals whose limb or finger has been lost. In one study, psychologists used magnetic source imaging (MSI) to measure neuronal activity in the brains of nine string players (six violinists, two cellists, and a guitar player) and six controls who had never played a musical instrument, when the thumb and fifth finger of the left hand were exposed to mild stimulation. The researchers felt that stringed instrument players, who use the fingers of their left hand extensively, might show different behavior in the brain – as a result of this extensive physical activity – than individuals who did not play stringed instruments.

The R code below reads in the data, which contains two variables:

- `Years` is the number of years the individual has played a stringed instrument (0 for people in the control group)
- `Activity` is a summary measure of neuronal activity from the MSI; a higher value indicates more neuronal activity.

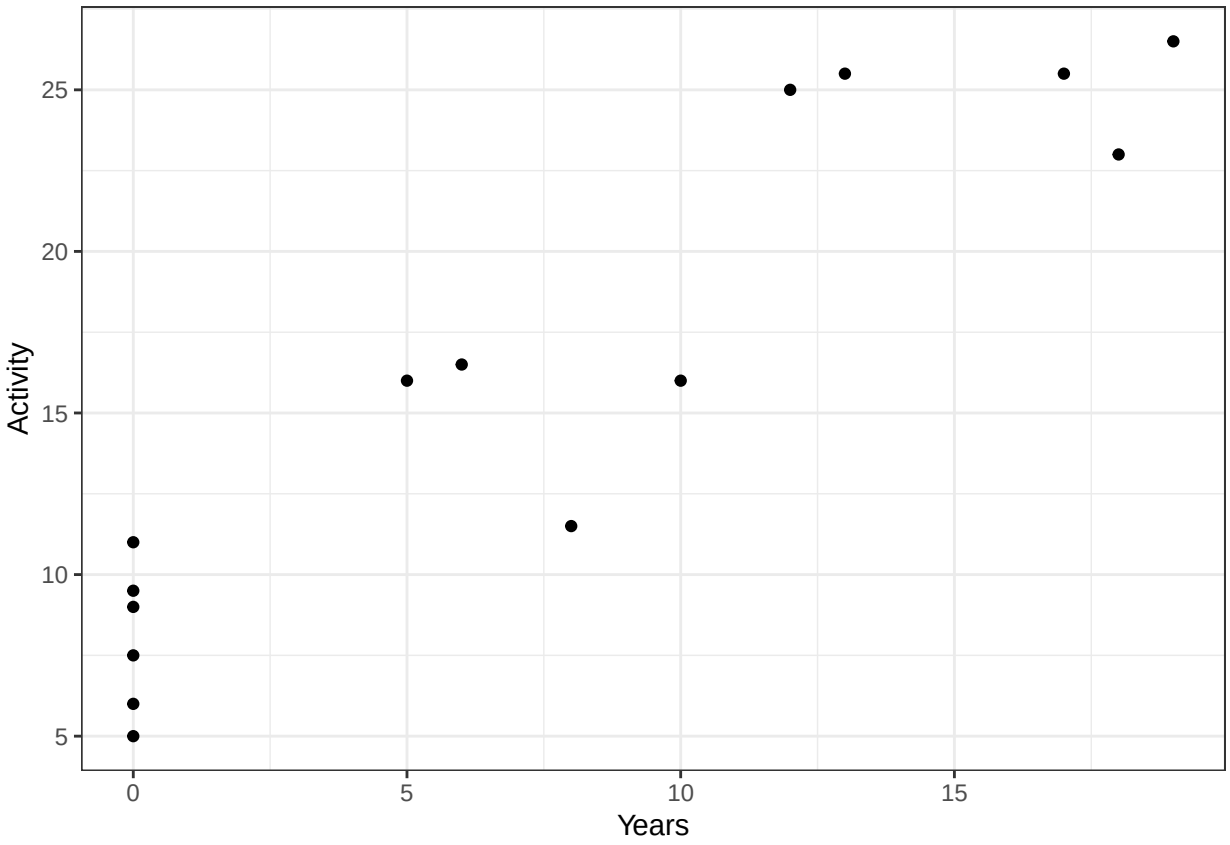
```
strings <- read.csv("http://www.evanlray.com/data/sleuth3/ex0728_string_instruments_brain_activity.csv")
```

(a) Is it possible to draw causal inferences about the relationship between the number of years someone has played a stringed instrument and their neuronal activity when the left hand is stimulated based on this study design? Explain.

No, this is an observational study, so we cannot draw causal inferences. People are not randomly assigned to play stringed instruments or not play stringed instruments. They also are not assigned randomly to playing string instruments for particular lengths of time (years). There could be other confounding variables that account for any observed association between years playing a stringed instrument and neuronal activity.

(b) Create a scatterplot of the data set with the explanatory variable on the horizontal axis and the response on the vertical axis.

```
ggplot(data=strings, aes(x=Years, y=Activity)) +
  geom_point() +
  theme_bw()
```



(c) What is the estimated equation of the line describing the relationship between the number of years someone has played a stringed instrument and their neuronal activity when the left hand is stimulated based on this data set?

```
lm_strings <- lm(Activity ~ Years, data=strings)
summary(lm_strings)
```

```
##
## Call:
## lm(formula = Activity ~ Years, data = strings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8644 -2.3730  0.1614  2.3713  4.6471
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.3873     1.1149   7.523 4.35e-06 ***
## Years         0.9971     0.1110   8.980 6.18e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.009 on 13 degrees of freedom
## Multiple R-squared:  0.8612, Adjusted R-squared:  0.8505
## F-statistic: 80.63 on 1 and 13 DF,  p-value: 6.178e-07
```

$$\hat{\mu}(Y|X = x) = 8.387 + 0.997x$$

(d) What are the interpretations of the estimated intercept and slope? Please interpret the coefficient estimates in context.

The estimated mean level of neuronal activity among people from a population similar to those enrolled in this study who never played a string instrument is 8.387.

(e) Find a 97.5% confidence interval for the intercept, β_0, by using the estimate and its standard error from the linear model fit summary output and results from an appropriate call to `qt`. If you want, you can also double check your answer with a call to `confint`. No need to interpret the interval in context or discuss.

Our interval has the form: $\hat{\beta}_0 \pm t^* SE(\hat{\beta}_0)$, where t^* is the critical value corresponding to the 97.5th percentile in a t distribution with 15-2=13 degrees of freedom.

```
## Find interval using qt()
t_crit <- qt(0.9875, df=13)
b0_hat <- summary(lm_strings)$coefficients[1,1]
se_b0_hat <- summary(lm_strings)$coefficients[1,2]
c(b0_hat-t_crit*se_b0_hat, b0_hat+t_crit*se_b0_hat)
```

```
## [1]  5.56365 11.21086
```

```
## Confirm with confint
confint(lm_strings, level=0.975)
```

```
##                 1.25 %   98.75 %
## (Intercept) 5.5636496 11.210860
## Years       0.7159027  1.278378
```

(f) Find the estimated mean level of neuronal activity for people who have been playing string instruments for 5 years. You should do this "by hand" using the estimated equation from part (c).

$$\hat{\mu}(Y|X = 5) = 8.387 + 0.997 \times 5 = 13.372$$

(g) Find a set of two Bonferroni adjusted confidence intervals for the mean level of neuronal activity for people who have never played a string instrument, and for people who have been playing string instruments for 5 years. Your intervals should have a familywise confidence level of 95%. You can use the `predict` function, no need to do this by hand. You do not need to discuss or interpret the intervals in this part.

```
## Create data frame to use with predict function
predict_data <- data.frame(
  Years = c(0, 5)
)

## Find confidence interval
predict(lm_strings, newdata = predict_data, interval="confidence", level=0.975)
```

```
##         fit      lwr      upr
## 1  8.387255  5.56365 11.21086
## 2 13.372958 11.31012 15.43579
```

(h) Interpret your intervals from part (g) as a family. As part of your answer, include a description of the phrase "95% confident" in the context of familywise intervals.

We are 95% confident that the mean neuronal activity level among people who have never played a string instrument is between 5.56 and 11.21, and that the mean neuronal activity level among people who have played a string instrument for 5 years is between 11.31 and 15.44. For 95% of samples, both of the intervals constructed in this way will simultaneously contain the means they are estimating.

(i) Explain in a sentence or two how your result from part (g) for the mean among people who have never played a string instrument relates to the confidence interval for the intercept from part (e).

Since the intercept of the model describes the mean level of neuronal activity among people who have never played a string instrument in the population, the confidence interval for part (g) is the same as the confidence interval for the intercept for part (e).

(j) Make another scatterplot of the data, this time including the estimated regression line and Scheffe-based 95% confidence bands.

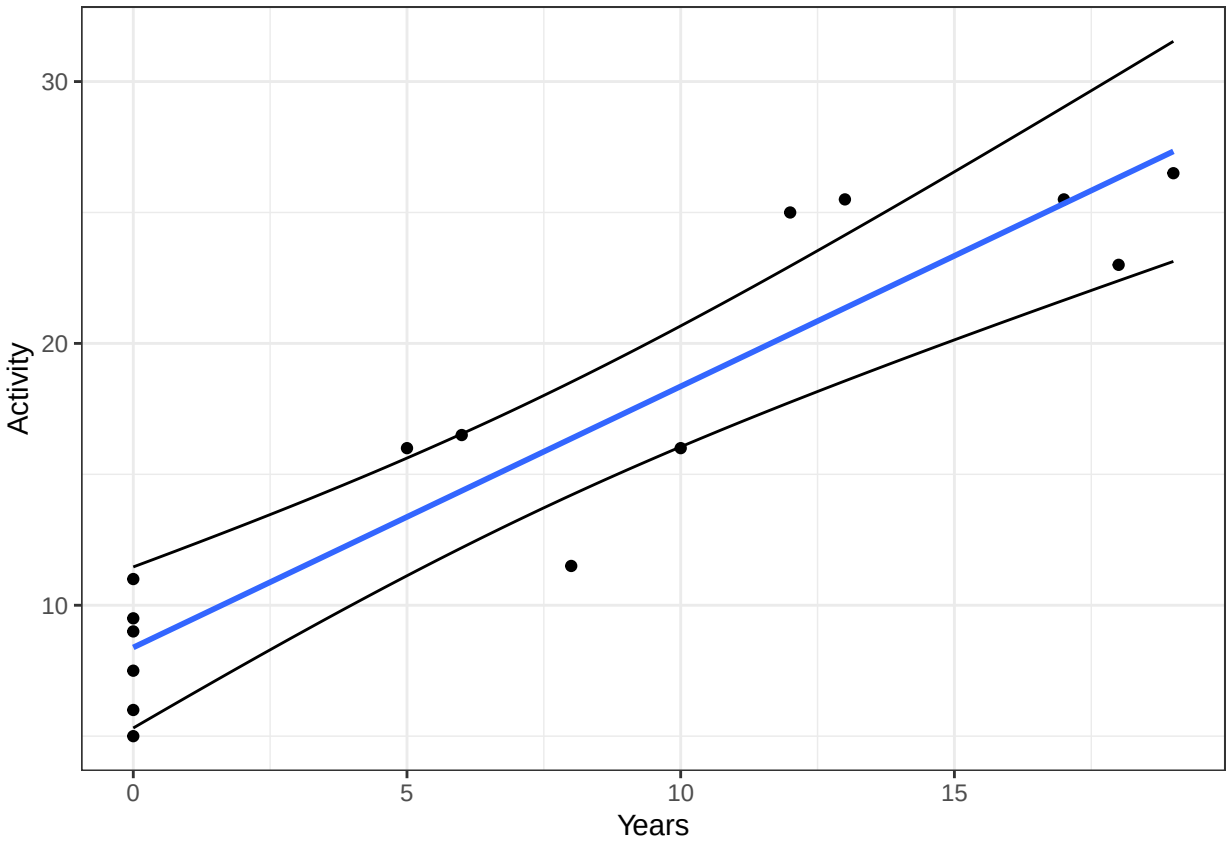
```r
library(lava) ## contains the scheffe function

## Create data frame to use with predict function
predict_df <- data.frame(
  Years=seq(from=0, to=19, length=100)
)

## Create Scheffe intervals
scheffe_cis <- scheffe(lm_strings, predict_df)

## Add intervals to predict_df for plotting purposes
predict_df <- predict_df %>% mutate(
  scheffe_lwr = scheffe_cis[,2],
  scheffe_upr = scheffe_cis[,3]
)

ggplot(data=strings, aes(x=Years, y=Activity)) +
  geom_point() +
  geom_smooth(method="lm", se=FALSE) +
  geom_line(data=predict_df, aes(x=Years, y=scheffe_lwr)) +
  geom_line(data=predict_df, aes(x=Years, y=scheffe_upr)) +
  theme_bw()
```



(k) Explain how to interpret the Scheffe-based confidence bands in part (j).

We are 95% confident that the population mean neuronal activity for people with a certain number of years of experience playing stringed instruments lies between the two black bands, at every value of number of years along the horizontal axis of the plot. Basically, we are 95% confident that the line describing the relationship between number of years playing a stringed instrument and mean neuronal activity is fully contained within the banded region (between the two black lines). For 95% of samples, the banded region obtained based on that sample would contain the population line.

(l) Conduct a hypothesis test of the claim that there is no association between the number of years someone has played a string instrument and their expected/mean level of neuronal activity as measured by the MSI. State your hypotheses using equations and a written sentence explaining the meaning of the hypotheses, and interpret your results in context.

$H_0 : \beta_1 = 0$; There is no linear association between the number of years someone has played a stringed instrument and their mean level of neuronal activity as measured by the MSI.

$H_A : \beta_1 \neq 0$; There is a linear association between the number of years someone has played a string instrument and their mean level of neuronal activity as measured by the MSI.

From the R summary output from part (c), the p-value for this test is 6.18×10^{-7}, so there is very strong evidence against the null and in favor of the alternative that there is a linear association between the number of years someone has played a stringed instrument and their mean level of neuronal activity as measured by the MSI.