HW8 - Stat 139

Callin Switzer

November 3, 2014

#1a

It would be appropriate to drop the intercept when you know the intercept is equal to zero.

#1b

Sons of very tall fathers will tend to be shorter than their fathers because of regression toward the mean. This happens because father height is based on genes and other random influences – such randomness can cause extreme heights. The random variance in the fathers' heights doesn't affect the random variance in the sons' heights. Thus, sons of really tall fathers tend to be shorter than their fathers, but their genes still make them relatively tall.

#1ca

The linear model assumes whether there is a linear relationship between stress and volume. However, an ANOVA test simply determines if there are any differences in the means among groups.

#1cb

In the linear model, there are three parameters: β_1, β_0, and σ. In ANOVA, we are estimating 10 parameters: μ_1, μ_2, ... , μ_9 and σ.

#1cc

For the linear model, the df = 45 - 2 = 43. For the ANOVA, df = 45 - 9 = 36.

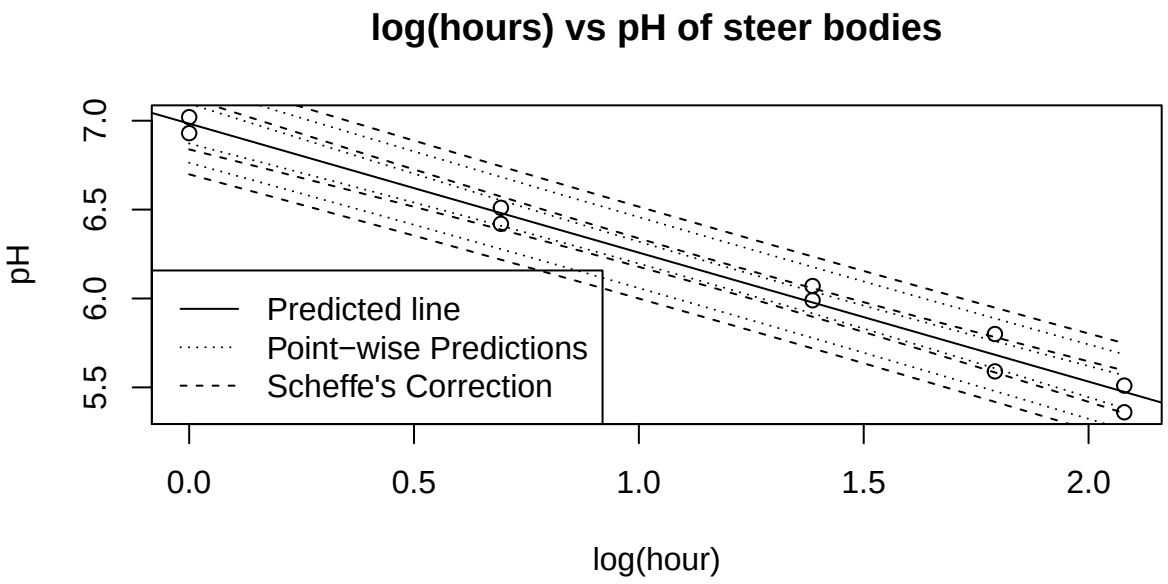
#2iii

Below, I've used the trick of subtracting log(5) from X to make the intercept at X = log(5)

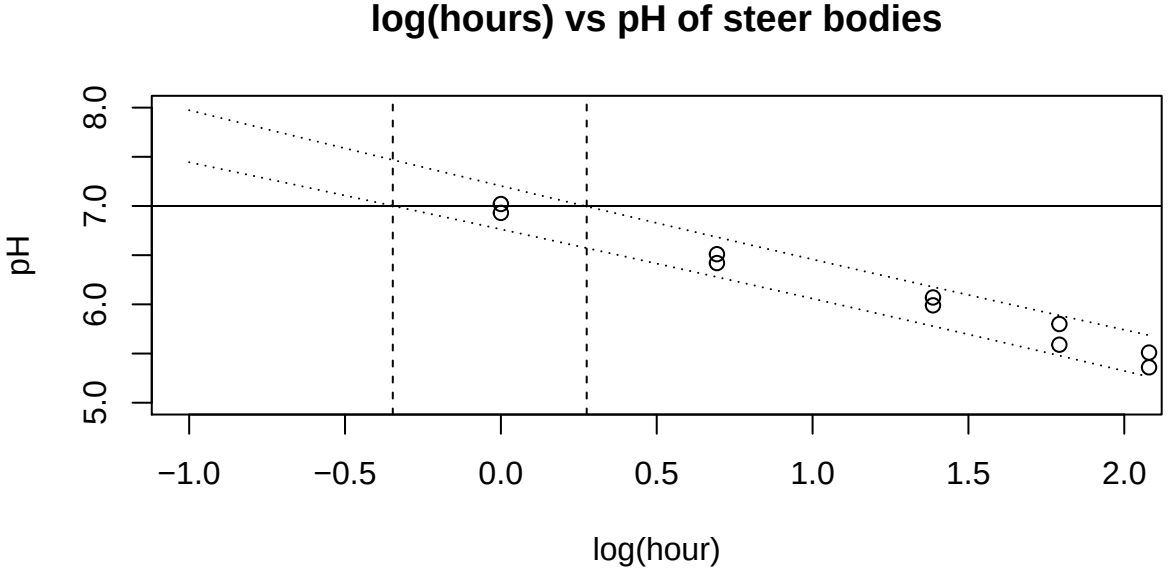
From the R output, you can see that the estimate is 5.816, and SE is 0.029, which is what I calculated by hand.

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  5.8157249 0.02974967 195.48869 5.247079e-16
## x           -0.7256578 0.03442633 -21.07857 2.695158e-08
```

#2v



#2vi



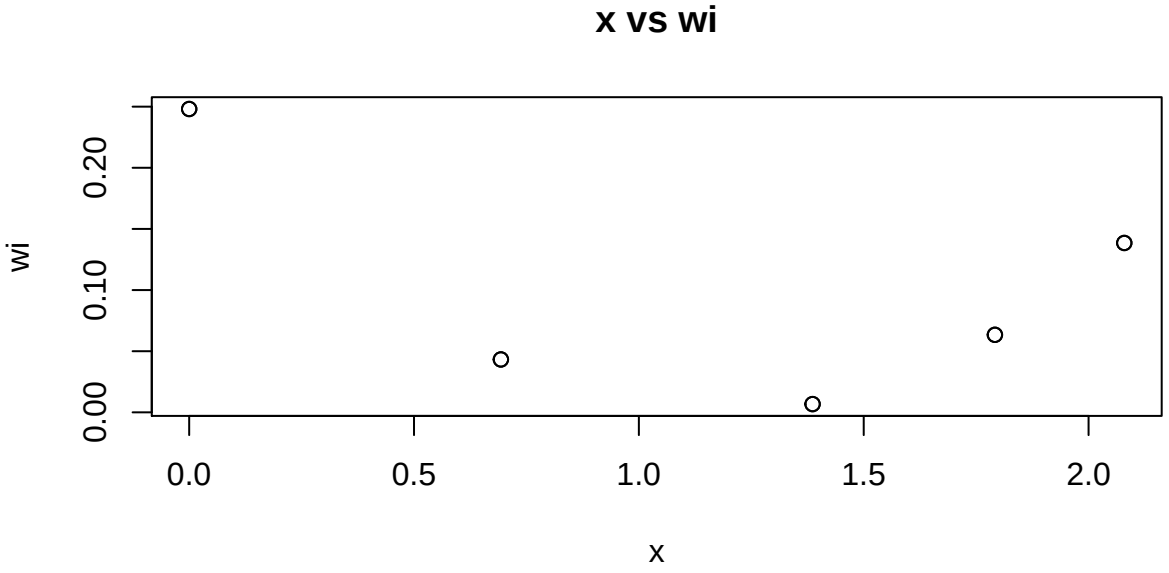
I used the graph with prediction intervals (without Scheffe's correction, because the example in the book doesn't use them). I estimated the calibration interval graphically. The calibration interval is -0.3467851 to 0.2753243 log(hours). When converted from the log values, it would be 0.7069572 to 1.3169577 hours

#2vii

We interpret w_i is proportional to the leverage. It is a measure of how much each point affects the slope of the line. As you can see from the plot below, as xi gets further from $\bar{x}$, w_i gets larger. The intuition behind

this pattern is that the points further from the mean have higher leverage, because their residuals will change more as the slope changes. The points that affect the slope the most are the points furthest from $\bar{X}$

We interpret $\left(\frac{Y_i-\bar{Y}}{X_i-\bar{X}}\right)$ as estimates of the slope (i.e. change in y divided by change in x). These are weighted by w_i.

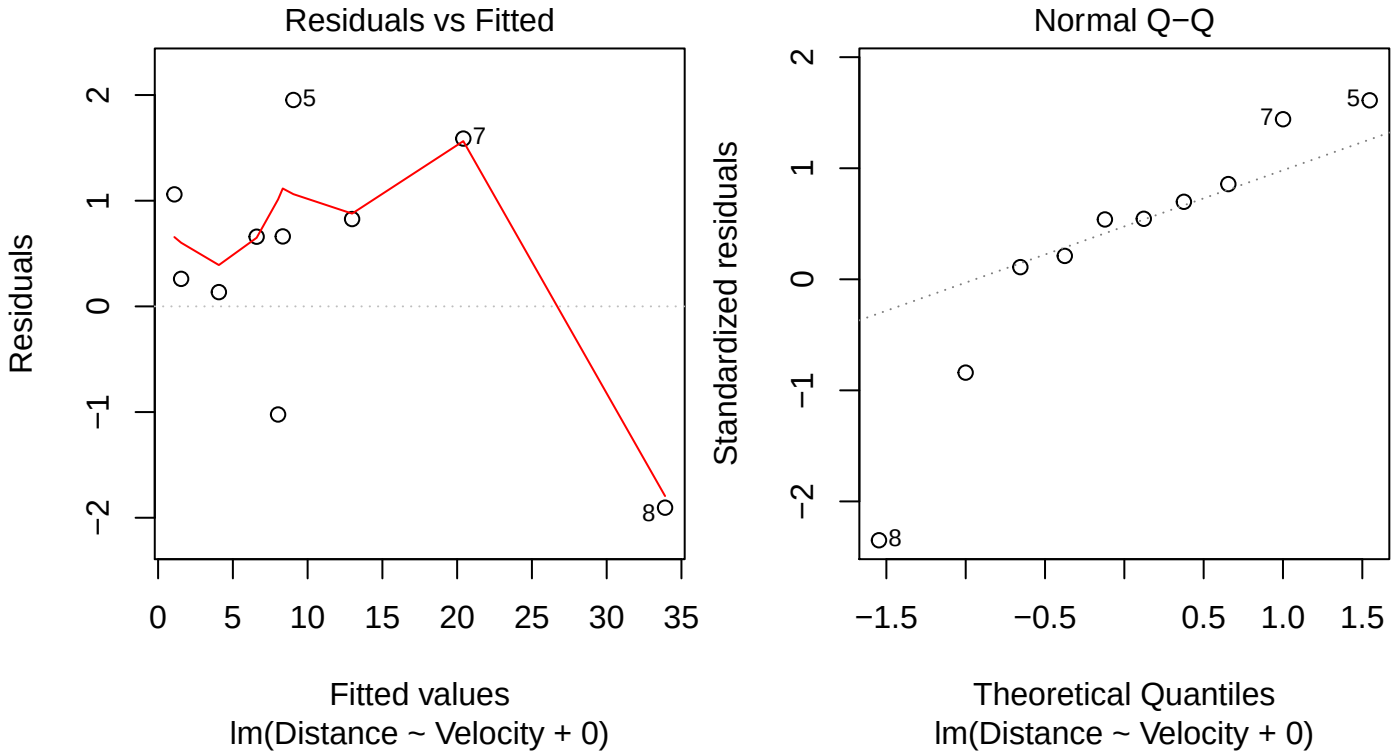


#3

The data are consistent with the hypothesis that the intercept = 0.

The estimated age of the universe is the slope * 979.8, or 1.6098098 billion years.

The linearity assumption is probably plausible, though it is difficult to tell with such few data points. I looked at the normal Q-Q plot and residual plot of the data, and the residuals do not quite look like they are normally distributed. All of the residuals besides two are above the 0 line. I am a little concerned about the equal variance assumption because the residuals look like they might be "fan-shaped" a little bit. The Q-Q plot shows that the residuals may not be normally distributed.



Statistical Report

Summary:

If the theory is taken as correct, then the estimated age of the universe is 1.6098098 billion years. This is the estimate of the slope when the line goes through the origin. A 95% confidence interval for the age is 1.589 to 1.801 billion years. This confidence interval is consistent with the confidence interval calculated in case study 7.1.1. The estimated slope in this case is slightly lower than case study 7.1.1.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.953678	0.525094	1.816204	0.10687
Velocity	0.001643	0.000064	25.668417	0.00000

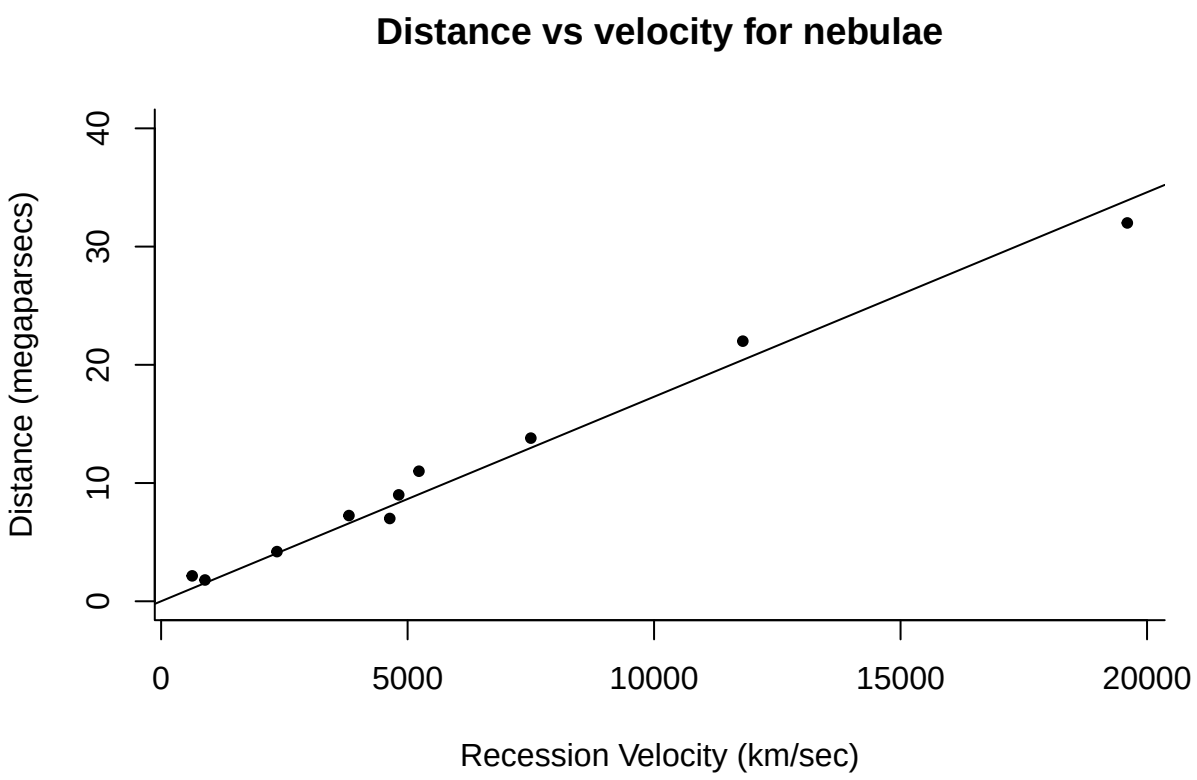
Table 1: Model fitted with intercept

	Estimate	Std. Error	t value	Pr(>|t|)
Velocity	0.001730	0.000048	36.270797	0.00000

Table 2: Model fitted without intercept

The data are consistent with the Big Bang Theory as proposed because the mean measured distance vs velocity is approximately a straight line, and the value of the line at zero velocity is zero. The result of a test for an intercept of 0 has a p-value of 0.106 (Table 1). Because this intercept was not not estimated to be different from zero, the model was refit with the intercept equal to zero (Table 2). The newly fit model was used to estimate the slope. This model is not consistent with the more wildely accepted age of the universe being between 10 and 15 billion years old.

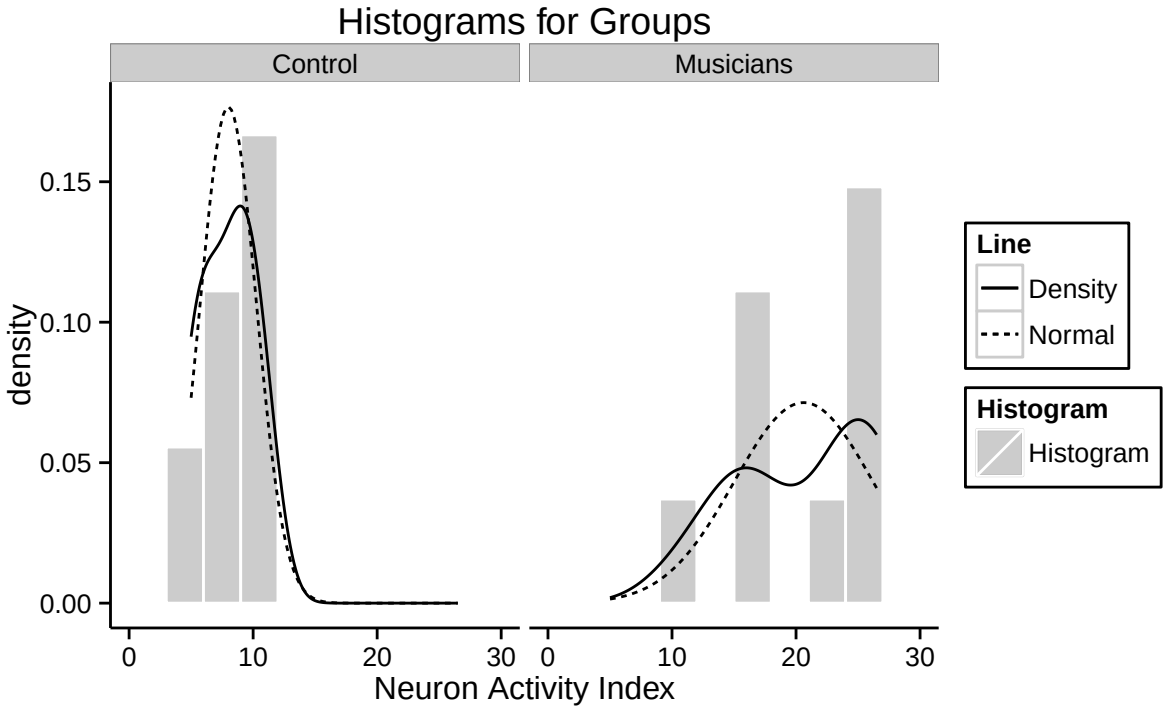
Graphical Display



Details for methods:

I used a linear regression to find the slope and intercept of the line for distance vs velocity. Because I failed to reject the hypothesis that the intercept = 0, I dropped that term from the model. I used a newly fit model with intercept = 0 to estimate the slope (see Table 1 and 2).

#4a



I used the histograms above to help evaluate the assumptions of the t-test. The t-test assumptions are the following: Independence between units, * Independent within populations, * Independent between populations, Homogeneity of units within each population, * equal means and variances within each population, * equal variances between populations, Populations are normally distributed, Random sampling from populations

The assumptions that may not be met are homogeneity of units. The musicians may not be homogeneous because they have been playing for different numbers of years. The variances between populations are not the same. The ratio of variances is 0.163272566 which is smaller than 0.5. This suggests that we might want to use a t-test with unpooled variance. However, the t-tools are fairly robust to unequal variance. The populations may not be normally distributed, but it is difficult to tell with such a small sample.

Below is the result of the two-sample t-test with pooled variance.

```
##              t         pvalue
## 5.198799538656 0.000171448424
```

We reject the hypothesis that the means (of Neuron activity) for musicians and non-musicians is equal.

#4b

Below is the result from a linear model comparing activity in musicians and non musicians. The t-value and p-value for the hypothesis test where H0 is that the slope = 0 is shown below. These values are equal to the values of the t-test with pooled variance that I used in #4a.

```
##              t         pvalue
## 5.198799538656 0.000171448424
```

We reject that hypothesis that the slope = 0.

#4c

The test statistics and p-values from #4a and #4b are the same. β_1 in (b) is 12.611111111 and the difference in sample means in (a) is 12.611111111. These are the same.

#4d

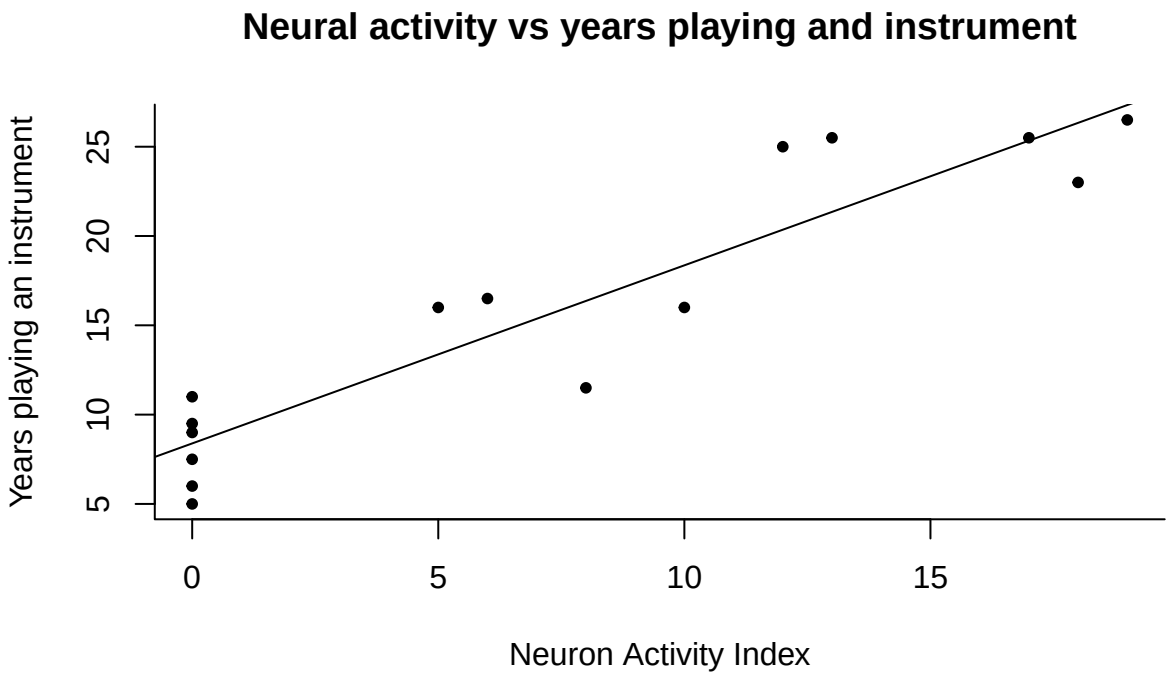
I addressed the question using a linear regression.

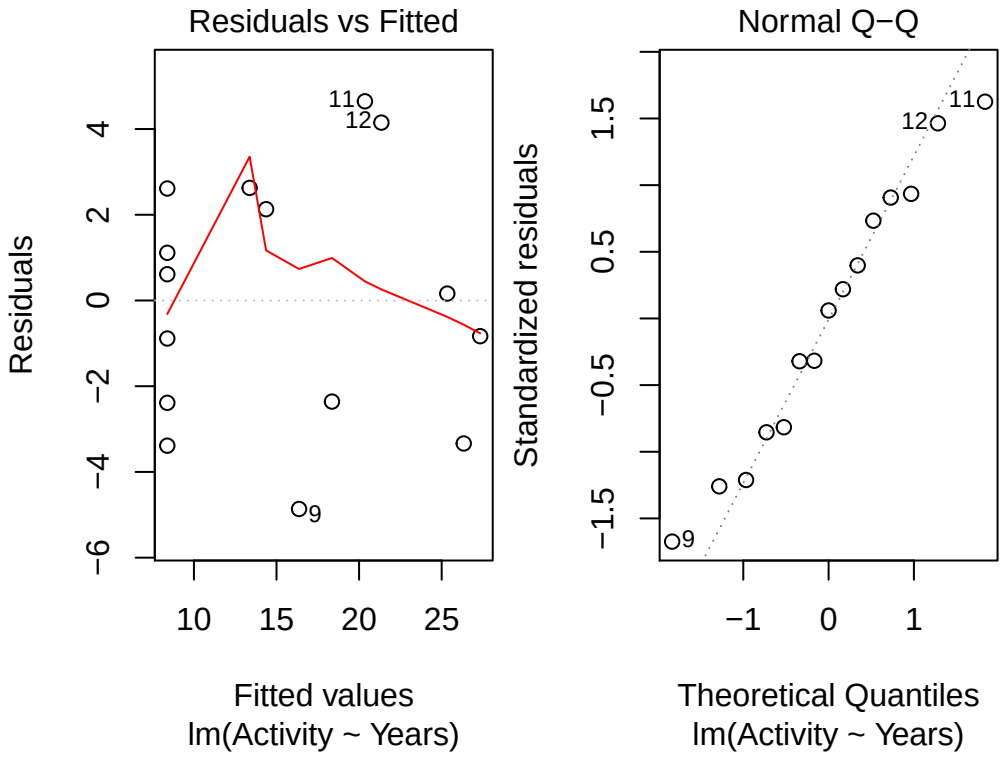
The assumptions of linear regression for this problem are the following: * There is a normally distributed subpopulation of responses (activity of neurons) for each value of the explanatory variable (the number of years playing an instrument). * The means of the subpopulations fall on a straight line, as a function of the number of years the person has been playing. * The variance of the residuals is equal along the whole length of the line. * The selection of an obersvation from any group is independent of teh selection of any other observation – meaning that knowing information about one person won't give you information about another person in that same group.

The assumptions look plausable. The residual plot shows that the residuals might not have equal variance, but it's hard to tell with a small sample. The normal Q-Q plot looks like the residuals are fairly normally distributed. The other assumptions seem to be plausible, based on the study design.

Statistical findings

We're interested in the slope of the linear regression. Below is a graph of the data, with a regression line and the residual plot and Q-Q plot





Here is a summary table of the linear regression. Our H_0 for the slope is that $\beta_1 = 0$. We reject the null hypothesis. From table 3, you can see that the p-value associated with the t-test for this hypothesis is much less than 0.05 (See row 2 of Table 3)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	8.387255	1.114887	7.522963	0.000004
Years	0.997141	0.111045	8.979574	0.000001

Table 3: Summary of linear regression for Neuron activity vs Number of years playing an instrument

#4e

I would interpret the estimate of the slope as the average estimated change in neural activity for each additional year of playing an instrument. For instance, I would say that for each year that people play an instrument, there is an average increase in their neuron activity index by about 0.99.

I would interpret the intercept as the estimated mean neoron activity index for people with 0 years of playing a musical instrument.

Code

```
################# 1
steer <- 1:10
timeAfterSlaughter <- c(1,1,2,2,4,4,6,6,8,8)
pH <- c(7.02, 6.93, 6.42, 6.51, 6.07, 5.99, 5.59, 5.80, 5.51, 5.36)
x <- log(timeAfterSlaughter)
y <- pH
plot(x, y)
cor(x, y)
sd(x)
sd(y)
b1 <- cor(x, y) * sd(y)/sd(x)
mean(x)
mean(y)
b0 <- mean(y) - b1*mean(x)
b0 + b1 * log(5)
mod1 <- lm(y~x)
sigmaHat <- sqrt(sum(mod1$residuals^2)/(length(x) - 2))
sigmaHat * sqrt(1/10 + (log(5) - mean(x))^2/(9*var(x)))
sigmaHat * sqrt(1/10 + (1.609 -1.19)^2/(9*.63))
var(x)
mean(x) - log(5)
xst <- x - log(5)
mean(xst)
var(xst)
sigmaHat * sqrt(1/10 + (mean(xst))^2/(9*var(x)))
mean(y) - mean(xst)*b1
SEb0 <- sigmaHat * sqrt(1/10 + (mean(x))^2/(9*var(x)))
SEb0
b0/SEb0
summary(mod1)
2*(1-(pt(143.89, 8)))
sigmaHat*sqrt(1/(9*var(x)))
b1/(sigmaHat*sqrt(1/(9*var(x))))
2*(pt(-21.08, 8))
# prediction interval
sigmaHat * sqrt(1 + 1/10 + (log(5) - mean(x))^2/(9*var(x)))
5.816 + 2.306*0.0874
################## #2iii
steer <- 1:10
timeAfterSlaughter <- c(1,1,2,2,4,4,6,6,8,8)
pH <- c(7.02, 6.93, 6.42, 6.51, 6.07, 5.99, 5.59, 5.80, 5.51, 5.36)
x <- log(timeAfterSlaughter)
y <- pH
x <- x-log(5)
mod2 <- lm(y~x)
tricky <- summary(mod2)
tricky$coefficients
```

```
################## #2v
x <- log(timeAfterSlaughter)
mod1 <- lm(y~x)
plot(x, y, main = "log(hours) vs pH of steer bodies", xlab = "log(hour)",
     ylab = "pH")
abline(mod1)
# point-wise confidence interval
newx <- seq(from = min(x), max(x), length.out = 100)
#newx <- x
dat <- predict(mod1, newdata=data.frame(x=newx),
        interval = c("confidence"), type="response")
lines(newx, dat[,2], lty = 3); lines(newx, dat[,3], lty = 3)
# confidence bands correcting for multiple testing
prdScheffe <- predict(mod1, newdata=data.frame(x=newx),
        interval = c("confidence"), type="response")
means <- prdScheffe[,1]
lwrBound <- prdScheffe[,2]
                  # mean   # SE                        # scheffe correction
prdScheffe[,2] <- means - (lwrBound-means)/qt(0.975,8)*sqrt(2*qf(0.95,2,8))
prdScheffe[,3] <- means + (lwrBound-means)/qt(0.975,8)*sqrt(2*qf(0.95,2,8))
lines(newx,prdScheffe[,2], col="black", lty=2)
lines(newx,prdScheffe[,3], col="black", lty=2)
# point-wise prediction interval
newx <- seq(from = min(x), max(x), length.out = 100)
#newx <- x
dat <- predict(mod1, newdata=data.frame(x=newx),
        interval = c("predict"), type="response")
lines(newx, dat[,2], lty = 3); lines(newx, dat[,3], lty = 3)
# confidence bands correcting for multiple testing
prdScheffe <- predict(mod1, newdata=data.frame(x=newx),
        interval = c("prediction"), type="response")
means <- prdScheffe[,1]
lwrBound <- prdScheffe[,2]
sigmaHat <- sqrt(sum(mod1$residuals^2)/(length(x) - 2))
SE <- sigmaHat * sqrt(1/10 + (newx - mean(x))^2/(9*var(x)))
SEP <- sigmaHat * sqrt(1 + 1/10 + (newx - mean(x))^2/(9*var(x)))

# confidence bands without scheffe's correction - checking my calcs
#lines(newx, prdScheffe[,1] + SEP*qt(p = .975, df = 8), col = "red")
#lines(newx, prdScheffe[,1] + SEP*qt(p = .975, df = 8), col = "red")

# confidence bands with scheffe's correction -- checking my own calcs
#lines(newx, prdScheffe[,1] + SEP*sqrt(2*qf(0.95,2,8)), col = "red")
                  # mean   # SE                        # scheffe correction
prdScheffe[,2] <- means - (lwrBound-means)/qt(0.975,8)*sqrt(2*qf(0.95,2,8))
prdScheffe[,3] <- means + (lwrBound-means)/qt(0.975,8)*sqrt(2*qf(0.95,2,8))
lines(newx,prdScheffe[,2], col="black", lty=2)
lines(newx,prdScheffe[,3], col="black", lty=2)
legend("bottomleft",
       legend = c("Predicted line", "Point-wise Predictions", "Scheffe's Correction"), lty = c(1, 3,2))
################## #2vi
# point-wise prediction interval
plot(x, y, main = "log(hours) vs pH of steer bodies", xlab = "log(hour)",
```

```
     ylab = "pH", xlim = c(-1,2), ylim = c(5, 8))

newx <- seq(from = min(x) - 1, max(x), length.out = 100)
#newx <- x
dat <- predict(mod1, newdata=data.frame(x=newx),
        interval = c("prediction"), type="response")
lines(newx, dat[,2], lty = 3); lines(newx, dat[,3], lty = 3)
abline(h = 7)
abline(v = newx[round(dat[,2], 2) == 7.00], lty =2)
abline(v = newx[round(dat[,3], 2) == 7.00], lty =2)
################## #2vii
wi <- (x - mean(x))^2 / (sum((x - mean(x))^2))
plot(x, wi, main = "x vs wi")
#other <- (y-mean(y))/(x - mean(x))
#plot(x, other)
################## #3
options(xtable.comment = FALSE)
library(xtable)
bang <- read.csv("data/ex0725.csv")
mod3 <- lm(Distance~Velocity, data = bang)
s <- summary(mod3)
#s$coefficients
# fit model without an intercept
mod4 <- lm(Distance~Velocity + 0, data = bang)
s4 <- summary(mod4)
#s4$coefficients[1] * 979.8
# residual plot
plot(mod4, which = 1)
plot(mod4,which = 2)
options("scipen"=10, "digits"=9)
ff <- xtable(s, caption = "Model fitted with intercept")
digits(ff) <- c(6,6,6,6, 5)
print(ff, floating = T)
sf <- xtable(s4, caption = "Model fitted without intercept")
digits(sf) <- c(6,6,6,6,5)
print(sf, floating = T)
CI <- confint(mod4, parm = "bang$Velocity", level = 0.95)
CIB <- CI * 979.8
plot(bang$Distance~bang$Velocity, bty = "l",
     main = "Distance vs velocity for nebulae",
     xlab = "Recession Velocity (km/sec)",
     ylab = "Distance (megaparsecs)", pch = 20,
     ylim = c(0, 40))
abline(mod4)
################## #4a
brain <- read.csv("data/ex0728.csv")
trt <- ifelse(test = brain$Years == 0, yes = "Control", no = "Musician")
#boxplot(brain$Activity~trt)
require(ggplot2)
require(plyr)
tAssumpChecker <- function(group1,
                           group2,
                           group1title = "group1",
```

```
                          group2title = "group2",
                          graphTitle = "Histograms for Groups") {

    mdat <- data.frame(dat = c(group1, group2),
                       trt = c(rep(group1title, length(group1)),
                               rep(group2title, length(group2))))
    sp <- data.frame(row.names = 1:nrow(mdat))
    sp$PctChange <- mdat$dat
    sp$SpeedLimit <- mdat$trt

    grid <- with(sp, seq(min(PctChange), max(PctChange), length = 100))
    normaldens <- ddply(sp, "SpeedLimit",
                        function(df) {
                             data.frame(
                                       predicted = grid,
                                       density = dnorm(grid, mean(df$PctChange),
                                                       sd(df$PctChange)))
                        })

    # look at distributions of data
    ggplot(sp, aes(x = PctChange)) +
         # histogram
         geom_histogram(aes(y = ..density.., fill = "Histogram"), color = "white",
                        alpha = 0.2, binwidth = 3) +
         # kernel density line
         geom_line(aes(y = ..density..,  lty = "Density"), stat = 'density')+
         # normal line
         geom_line(aes(y = density, x = predicted, lty = "Normal"), data = normaldens) +
         # facet
         facet_grid(~SpeedLimit)  +

         # labels and theme
         #xlim(c(-3.5, 3.5))+
         labs(x = "Value", title = graphTitle) +
         theme_bw() +
         theme(legend.background = element_rect(colour = "black"),
              plot.background = element_blank()
              ,panel.grid.major = element_blank()
              ,panel.grid.minor = element_blank()
              ,panel.border = element_blank()
              ,axis.line = element_line(color = 'black')) +
         # Names for the legend
         scale_linetype(name = "Line")+
         scale_fill_manual(name = "Histogram", values = c("black"))
    ## the distribution looks like we need to log-transform the data
}
tAssumpChecker(group1 = brain$Activity[trt == "Control"],
               group2 = brain$Activity[trt == "Musician"],
               "Control", "Musicians") + xlab("Neuron Activity Index")
#var(brain$Activity[trt == "Control"])
#var(brain$Activity[trt == "Musician"])
tt <- t.test((brain$Activity[trt == "Musician"]),
             (brain$Activity[trt == "Control"])
```

```
       , var.equal = T)
c(tt$statistic, pvalue = tt$p.value)
################### #4b
brain$trt <- ifelse(brain$Years == 0, 0, 1)
#plot(brain$Activity~brain$trt)
mod6 <- lm(Activity~trt, data = brain)
#abline(mod6)
tm <- summary(mod6)
c(t = tm$coefficients[2,3], pvalue = tm$coefficients[2,4])
diffMeans <- mean(brain$Activity[trt == "Musician"]) - mean(brain$Activity[trt == "Control"])
################## #4c
mod6 <- lm(Activity~trt, data = brain)
xt <- xtable(mod6, caption = "Linear Model for Activity vs Musician/NonMusician")
digits(xt) <- c(7,7,7,7,7)
print(xt)
################ #4d
plot(brain$Activity~brain$Years, bty = "l",
     xlab = "Neuron Activity Index",
     ylab = "Years playing an instrument", pch = 20,
     main = "Neural activity vs years playing and instrument")
mod7 <- lm(Activity~Years, data = brain)
abline(mod7)
plot(mod7, which = 1)
plot(mod7, which = 2)
ta <- xtable(summary(mod7), caption="Summary of linear regression for Neuron activity vs Number of years
digits(ta) <- c(6,6,6,6,6)
print(ta)
```