Homework 4

STAT 4500

Cody Frisby

12/10/2016

5.6

	For	Against
For	41	27
Against	16	58

Running an exact test testing whether or not there was a change in attitude towards routinely asking patients about alcohol consumption before and after a video/discussion

$$H_0 : Before \text{ and } After \text{ are equal}$$

This is a binomial sign test with 27 minus signs and 16 plus signs, $n = 43$ and $p = 0.5$. Additionaly it is two-sided so

$$P(X \leq 16) = 2\sum_{x=0}^{16} \binom{43}{x} (0.5)^x (0.5)^{43-x} = 0.1262895$$

It appears there is not any significant evidence that the video/disscussion altered the attitudes of the general practitioners.

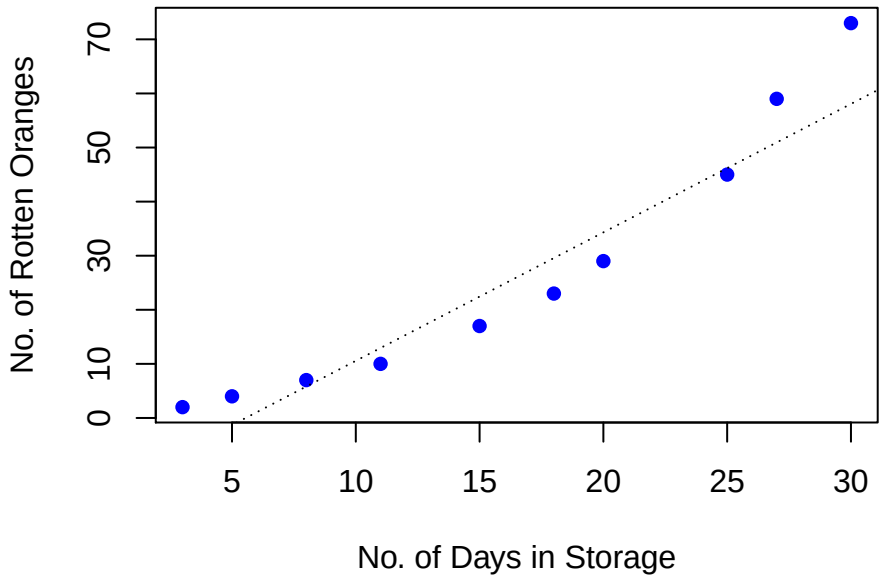
11.2

We get the residuals (e_i) but fitting a linear model (`lm` in `R`) and extracting the residuals. The sum of the residuals is 0. The sum the the residuals multiplied by the x's is also 0. The procudure of least squares estimation estimates β_0 and β_1 percisley by minimizing the differences of the fitted line and the observed data. Summing these differences, by definition, has to be zero.

11.3

A plot of the data is below with fitted Theil-Kendall line

$$\hat{y} = -13.1875 + 2.375x$$



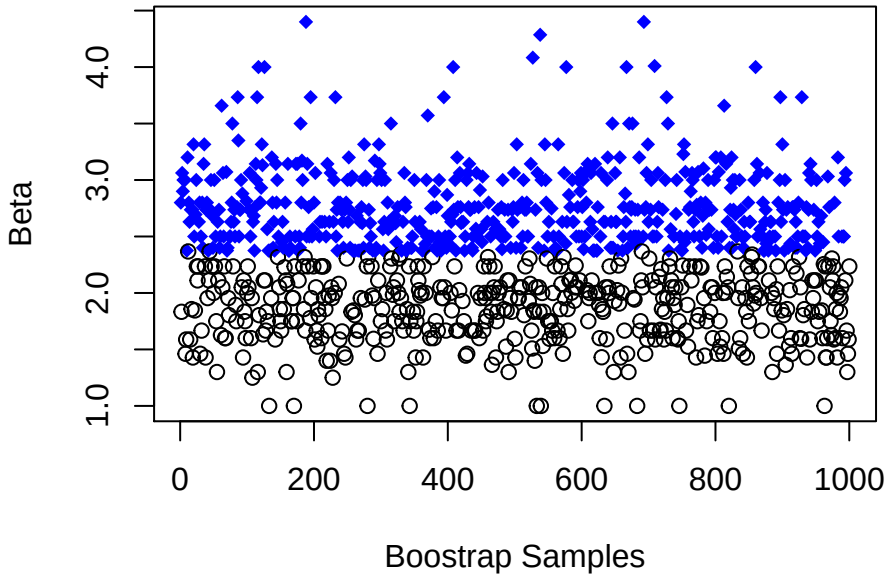
and the scatter plot of the data appears to show some non-linear behavior.

The confidence interval from bootstrapping apears to be much wider than the confidence interval based on the Theil-Kendall method.

$$[1.4, 3.5]$$

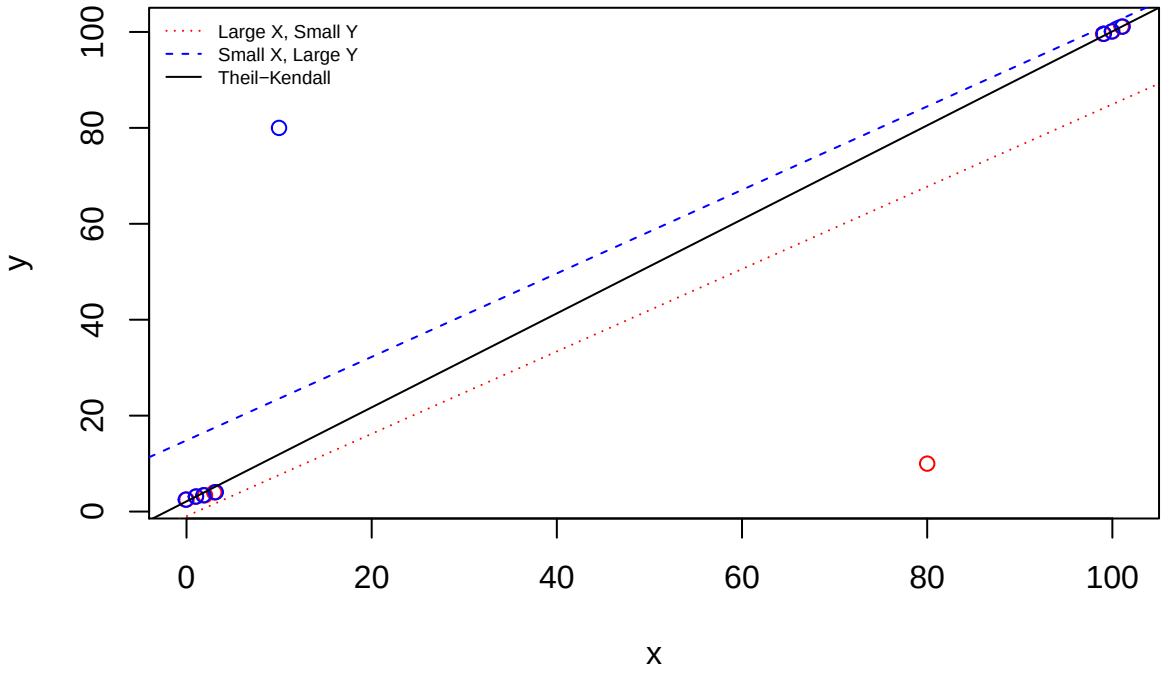
$$[2.034282, 2.821373]$$

Taking a look at a plot of all the bootstrapped betas with a layer of color where $\beta_{bootstrap} >= \hat{\beta}$ we can see that approximately 50% are greater than or equal to 2.375.



11.10

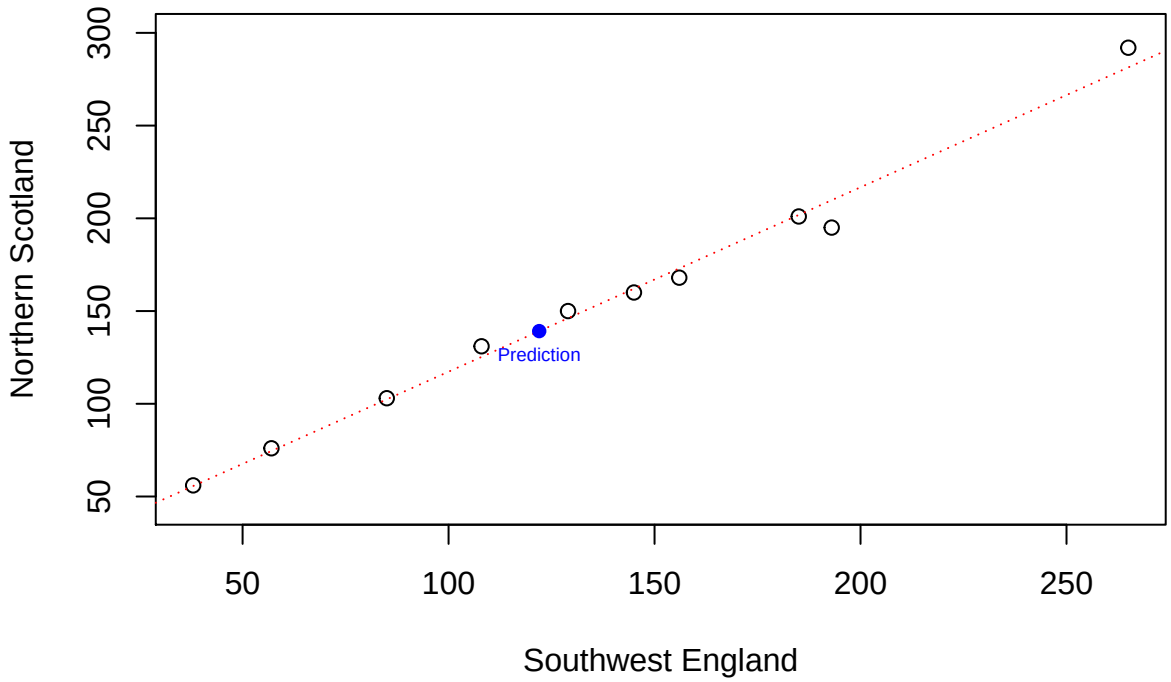
The estimate for β is 0.979798. A 95% confidence interval is $[0.7173274, 0.9837624]$. Below is a plot of the data with introduced outlier points.



Neither of these outliers have any influence on the fit of the line using the Theil-Kendall method. But they do influence the fit of the line using the least squares method as can be seen from the red and blue lines.

11.11

A scatter plot of the data appears to indicate a linear relationship between these two areas for the given plants. Ivy is somewhat seperated from the rest of the observations by quite a few days, but appears to be consistent with the linear relationship here. A linear model may be appropriate here.



Using

$$\hat{y} = 17.8609138 + 0.9944092x$$

where $\hat{y}$ would be the average days from January to the first flowering of a given plant from Southwest England (x) in North Scotland.

$$17.8609138 + 0.9944092(122) = 139.1788309$$

as can be seen on the above plot represented by the blue point.

12.9

	Excellent	Reasonable	Poor
UK	42	30	28
USA	20	41	19
France	19	29	12
Germany	26	22	12
Portugal	18	31	21
Brazil	31	42	7

Using a χ^2 test we find the test statistic $\chi^2 = 25.3400037$ with $(r-1)(c-1) = (5)(2) = 10$ degrees of freedom. There is evidence to suggest that the instructions are more acceptable in some countries than in others, $p = 0.004737$.

R Code:

```
#### 5.6 ####
a <- matrix(c(41, 16, 27, 58), nrow = 2, dimnames =
       list(After = c("For", "Against"), Before = c("For", "Against")))
rows <- apply(a, 1, sum)
cols <- apply(a, 2, sum)
test <- mcnemar.exact(a)
#### 11.2 ####
x <- 0:6
y <- c(2.5,3.1,3.4,4,4.6,5.1,11.1)
fit <- lm(y ~ x)
sse <- sum(fit$residuals)
xsse <- sum(x*fit$residuals)
#### 11.3 ####
x <- c(3,5,8,11,15,18,20,25,27,30)
y <- c(2,4,7,10,17,23,29,45,59,73)
fit <- mblm(y ~ x, repeated = F) # False for Theil-Kenkall method
b <- fit$coefficients
# Plot
plot(x,y, pch = 16, col = "blue", ylab = "No. of Rotten Oranges",
     xlab = "No. of Days in Storage")
abline(fit$coefficients[1], fit$coefficients[2], lty = 3)
# Bootstrap
B <- 1000
```

```
n <- 10
df <- data.frame(x, y)
boot <- numeric(B)
for(i in 1:B){
  s <- sample(1:n, n, replace = TRUE)
  df.samp <- df[s, ]
  boot[i] <- mblm(y ~ x, data = df.samp, repeated = F)$coef[2]
}
q.boot <- quantile(boot, c(0.025, 0.975))
q.ken <- confint.mblm(fit)[2,]
#### 11.10 ####
x <- c(0, 1, 2, 3, 99, 100, 101)
y <- c(2.5, 3.1, 3.4, 4, 99.6, 100.1, 101.1)
fit <- mblm(y ~ x, repeated = F)
conf.int <- confint.mblm(fit)
# Outlier point plots
x <- c(0, 1, 2, 3, 80, 99, 100, 101)
y <- c(2.5, 3.1, 3.4, 4, 10, 99.6, 100.1, 101.1)
fit <- mblm(y ~ x, repeated = F)
plot(y~x, col = "red", main = "")
abline(lm(y~x), lty = 3, col = "red")
x <- c(0, 1, 2, 3, 10, 99, 100, 101)
y <- c(2.5, 3.1, 3.4, 4, 80, 99.6, 100.1, 101.1)
fit <- mblm(y ~ x, repeated = F)
points(jitter(x),jitter(y), col = "blue")
abline(lm(y~x), lty = 2, col = "blue")
abline(fit, lty = 1)
# the line is the same using Theim-Kendall method.
legend("topleft",
       legend = c("Large X, Small Y", "Small X, Large Y", "Theil-Kendall"),
       bty="n", col=c("red", "blue", "black"), lty=c(3,2,1), cex = 0.6)
#### 11.11 ####
Sw.England <- c(38,57,85,108,129,145,156,185,193,265)
N.Scotland <- c(56,76,103,131,150,160,168,201,195,292)
df <- data.frame(Sw.England, N.Scotland)
df$plant <- c("Hazel", "Coltsfoot", "Wood.Anemone", "Hedge.Garlic",
              "Hawthorn", "White.Ox.Eye", "Dog.Rose", "Greater.Bindweed",
              "Harebell", "Ivy")
plot(Sw.England, N.Scotland, ylim = c(45, 300), xlab = "Southwest England",
     ylab = "Northern Scotland")
#text(Sw.England, N.Scotland, df$plant, cex = 0.6, pos = 1)
fit <- lm(N.Scotland ~ Sw.England)
# using a linear model we predict...
b <- fit$coefficients
p <- predict(fit, newdata = data.frame(Sw.England = 122))
x <- 122; y <- p
abline(fit, lty = 3, col = "red")
points(x,y, pch = 16, col = "blue")
text(x,y, "Prediction", pos = 1, cex = 0.6, col = "blue")
#### 12.9 ####
a <- matrix(c(42,20,19,26,18,31,30,41,29,22,31,42,28,19,12,12,21,7),
            ncol = 3)
colnames(a) <- c("Excellent", "Reasonable", "Poor")
```

```r
rownames(a) <- c("UK", "USA", "France", "Germany", "Portugal", "Brazil")
# we want to test if there is any indications of difference in views
rows <- apply(a, 1, sum)
cols <- apply(a, 2, sum)
N <- sum(a)
m <- outer(rows, cols, "*")/N
stat <- sum((m - a)^2/m)
test <- chisq.test(a, correct = FALSE)
```