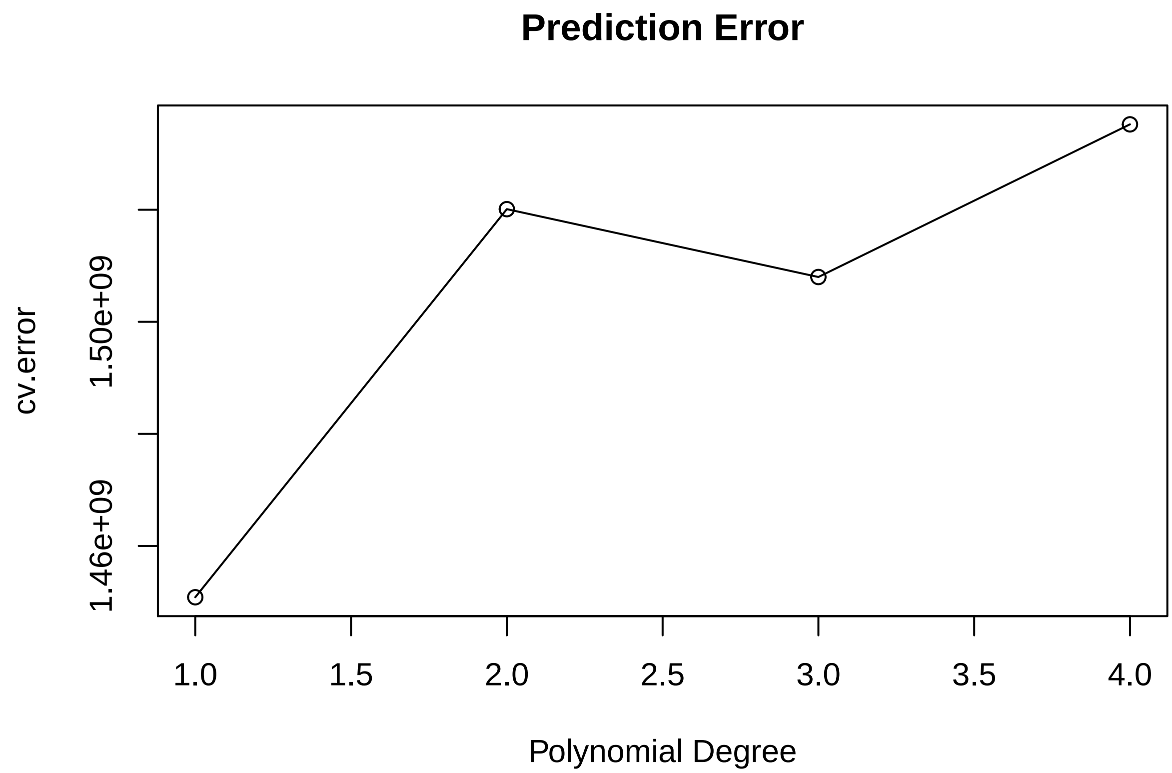
HW 3

Ben Buzzee

9/20/2019

1. Cross Validation

```r
library(boot)
library(MASS)
income <- read.csv("income.txt", sep = "", header = T)

glm.fit1 = glm(y~x, data=income)
cv.error1=cv.glm(income, glm.fit1, K=5)


glm.fit2 = glm(y~poly(x,2), data=income)
# quadratic model
cv.error2=cv.glm(income, glm.fit2, K=5)

glm.fit3 = glm(y~poly(x,3), data=income)
# cubic model
cv.error3=cv.glm(income, glm.fit3, K=5)

glm.fit4 = glm(y~poly(x,4), data=income)
# quartic model
cv.error4=cv.glm(income, glm.fit4, K=5)


cv.error=c(cv.error1$delta[1],cv.error2$delta[1],cv.error3$delta[1],cv.error4$delta[1])

d=c(1,2,3,4)
plot(d,cv.error, main = "Prediction Error", xlab = "Polynomial Degree")
lines(d,cv.error)
```

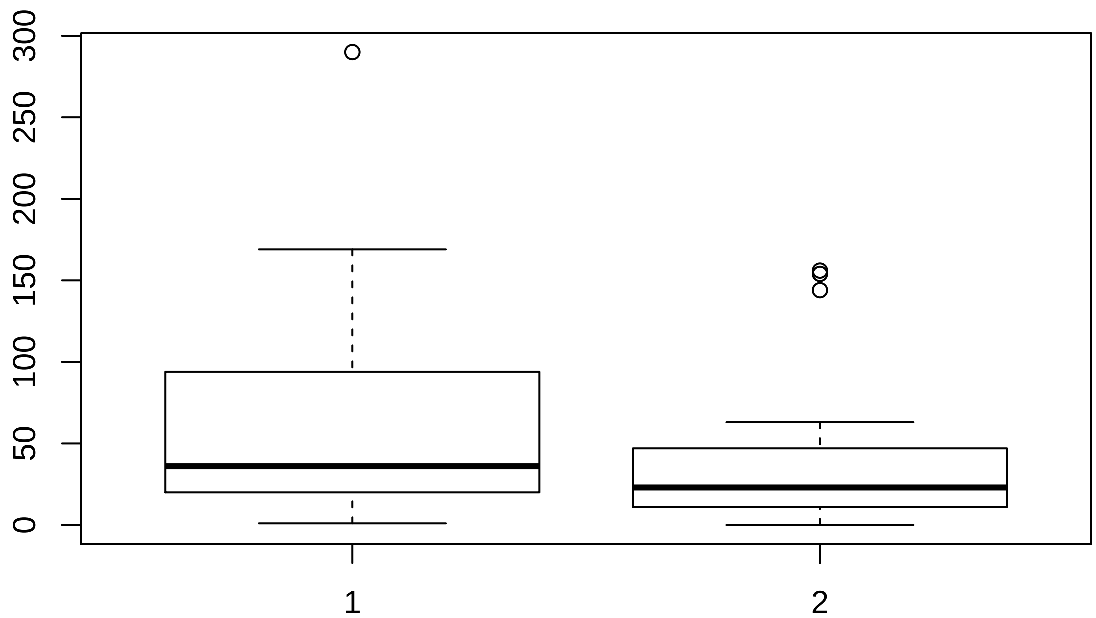


Depending on the situation a degree of either 1 or 4 would be acceptable. A linear model would be much more interpretable, while a polynomial of degree 4 has a smaller prediction error.

2 Mann-Whitney Test

We want to test the hypothesis that children exposed to violent TV were slower to react to real-life violence than children who were not exposed to violent TV. The response times were measured in seconds for both groups. First let's examine the data visually:

```
violent <- read.csv("violent.txt", sep = " ")
x <- violent$karate
y <- violent$olympics

boxplot(x,y)
```



(a)

If we let F be the CDF of violent tv watcher's reaction times, and G be the cdf of reaction times of children not exposed the violent TV, our hypothesis statements would be $H_0 : F(t) = G(t)$ vs $H_0 : F(t) = G(t - \Delta)$ for some $\Delta < 0$

```
wilcox.test(x,y, alternative = "greater")
```

```
##
##  Wilcoxon rank sum test
##
## data:  x and y
## W = 274, p-value = 0.09222
## alternative hypothesis: true location shift is greater than 0
```

With a U test statistic of 274 and a p-value of .09, would fail to reject the null hypothesis. However, I think that would be a misleading conclusion based on an arbitrarily chosen "significance" level and that a p-value of .09 suggest there is evidence that reaction times were longer for children exposed to violent TV.

To convert our U statistic to the W statistic, we can subtract $\frac{n(n+1)}{2}$ which in this case is 231 so our W statistic would be 43.

To conduct a hypothesis test based on a large sample approximation, we can set the exact argument to false in the wilcox.test function.

```
wilcox.test(x,y, alternative = "greater", exact = FALSE)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  x and y
## W = 274, p-value = 0.09122
## alternative hypothesis: true location shift is greater than 0
```

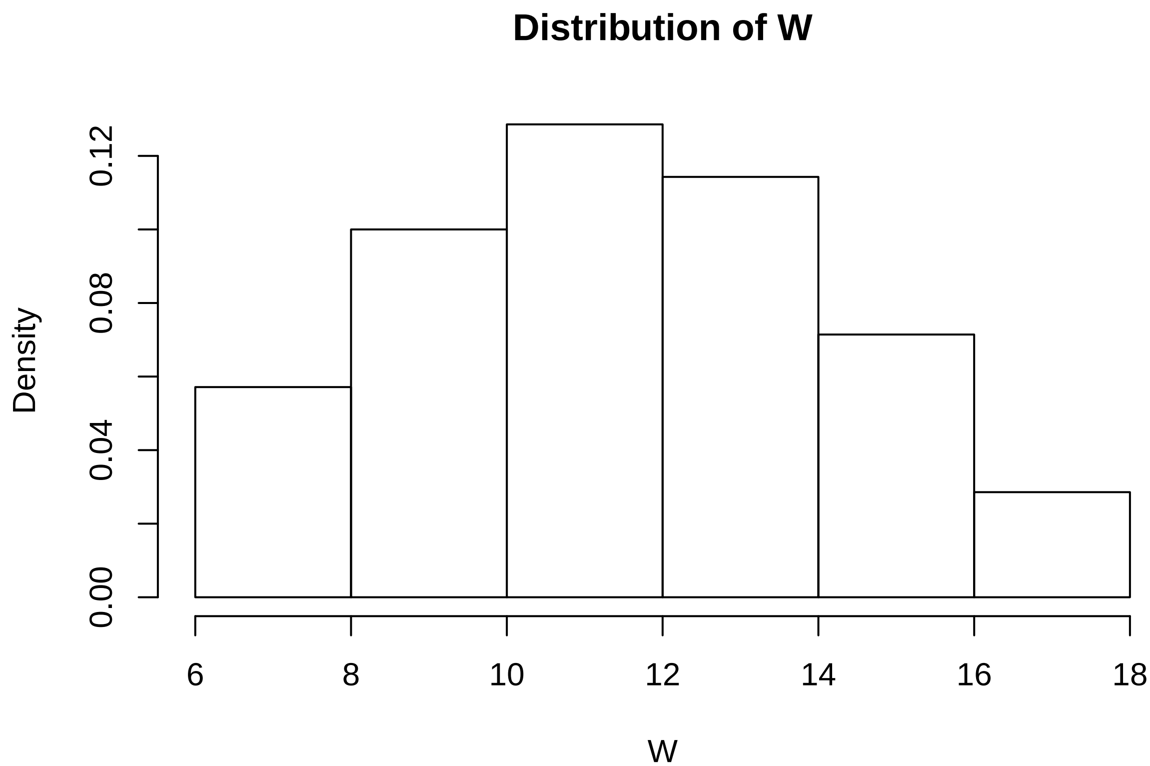
And we come to the same conclusion as before with our p-value about .001 lower.

4.

```
x=sort(c(2.1, 1.9, 2.6, 3.3))
y=sort(c(1.9, 2.6, 3.7, NA))
ranky <- c(1.5,4.5,7)

df  <- data.frame(x = x, y = c(y, NA), ranky = c(ranky, NA))
```

We can find the W statistic by summing the ranks of Y. This gives us a W statistic of 13.

```
x=sort(c(2.1, 1.9, 2.6, 3.3))
#remove NA
y=sort(c(1.9, 2.6, 3.7))

xy <- c(x,y)

ranksums <- colSums(combn(1:7, 3))


hist(ranksums, probability = T, main = "Distribution of W", xlab = "W")
```



To find how extreme W is, we can look at a tabled version of the above distribution:

```r
table(ranksums)
```

```
## ranksums
##  6  7  8  9 10 11 12 13 14 15 16 17 18
##  1  1  2  3  4  4  5  4  4  3  2  1  1
```

Our p-value would be the number of ways we can get a test statistic as or larger than 13 under the null hypothesis. This would be 15/35 or .43. So we would fail to reject the null hypothesis and have no evidence of a difference in the location of the distributions for X and Y.