Econometrics Homework 7

Timothy Schwieg

April 5, 2018

1 Question 1

1.1 a

He could simply drive the same distance back and forth many times, that is, drive the distance he walked n times in a row. There will still be the same error on his starting odometer, and error on his ending odometer, but since the value he is after is nd, his estimate of d which is $\frac{nd}{n}$ has all the errors divided by n. By using this strategy he may drive the error down to however small he wishes, at the cost of the time spent driving that distance.

1.2 b

Our measurement equation is given by:

$$\begin{aligned} S + nd + \xi_s &= E + \xi_e \\ nd &= E - S + \xi_e - \xi_s \\ d &= \frac{E - S + \xi_e - \xi_s}{n} \end{aligned}$$

where $\xi_i \sim U(0,1)$

Our estimator $\hat{d}$ is given by $\frac{E-S}{n} = d + \frac{\xi_s - \xi_e}{n}$

Since it is known that the mean-squared error is the bias squared plus the variance, we need only calculate the bias and variance of $\hat{d}$.

$$\mathbb{E}[\hat{d}] = d + \frac{\mathbb{E}[\xi_s] - \mathbb{E}[\xi_e]}{n} = d$$
$$\mathbb{V}(\hat{d}) = \frac{1}{n^2}\mathbb{V}(\xi_s - \xi_e) = \frac{1}{n^2}\left(\mathbb{V}(\xi_s) + \mathbb{V}(\xi_e)\right)$$
$$\frac{1}{n^2}\left(\frac{1}{12} + \frac{1}{12}\right) = \frac{1}{6n^2}$$

Thus the mean-squared error is given by: $\frac{1}{6n^2}$

1.3 c

Following the notion on page 69 of Foundations of Empirical Intuition by Harry Paarsch, it is known that the distribution of the difference between two uniform random variables is given by:

$$F_{\text{x}}(x) = \int_0^1 \int_{\text{-1}}^{\text{x}} 1 dx dy = \begin{cases} \frac{1}{2} + x + \frac{x^2}{2} & -1 \leq x \leq 0 \\ \frac{1}{2} + x - \frac{x^2}{2} & 0 \leq x \leq 1 \end{cases}$$

The density of the distribution is given by the derivative and is thus:

$$f_x(x) = \begin{cases} 1 + x & -1 \leq x \leq 0 \\ 1 - x & 0 \leq x \leq 1 \end{cases}$$

Now for a series of random samples on n days, we can note that

$$S + d + \xi_s = E + \xi_e$$
$$E - S = d + \xi_s - \xi_e$$

Thus the distribution of E-S is the difference of two uniforms, shifted over to the right by d. Thus the distribution of the difference is given by:

$$F_Y(y) = P(d + \xi_s - \xi_e \leq y) = P(\xi_s - \xi_e \leq y - d)$$
$$F_x(y - d) = \begin{cases} \frac{1}{2} + (y - d) + \frac{(y-d)^2}{2} & d - 1 \leq y \leq d \\ \frac{1}{2} + (y - d) - \frac{(y-d)^2}{2} & d \leq x \leq d + 1 \end{cases}$$

and its density is given by:

$$f_x(x) = \begin{cases} 1 + y - d & d - 1 \leq y \leq d \\ 1 - y - d & d \leq y \leq d + 1 \end{cases}$$

The mean and variance of the measurement error are given by:

$$\mathbb{E}[\xi_s - \xi_e] = .5 - .5 = 0$$
$$\mathbb{V}(\xi_s - \xi_e) = \mathbb{V}(\xi_s) + \mathbb{V}(\xi_e) = \frac{1}{6}$$

1.4 d

Since Y is given by the truth plus an error which has mean zero, its bias will be zero, and by the linearity of expected value, any linear combination of Y will be unbiased as well.

$$\mathbb{V}(\bar{Y}) = \mathbb{V}(\frac{1}{N}\sum_{n=1}^{N} y_n) = \frac{1}{N^2}\sum_{n=1}^{N}\mathbb{V}(y_n) = \frac{1}{6N}$$

By applying the Lindinberg-Levy Central Limit theorem, we may note that $\sqrt{N}\frac{\bar{Y}-d}{\sqrt{\mathbb{V}(Y)}} \sim N(0,1)$ Since $\bar{Y}$ is constructed by linear combinations of uniform distributions, this will converge in distribution to the normal quite quickly.

1.5 e

Since the variance of the sample average is convergign on an order of $O(\frac{1}{N})$ it is converging relatively slower than part a which is converging at a rate of $O(\frac{1}{N^2})$. This means that you need far fewer measurements, which makes intuitive sense, as the method in part A only has the two errors which are uniform(0,1) present, while the method suggested in the latter part has 2N errors which are all uniform(0,1).

2 Question 2

```
library( gtools )
library( ROCR )

set.seed( 235711 )

passFailData <- read.table( "PassFail.dat", header=FALSE )
```

```r
nObs <- nrow( passFailData )

#There was some issues with sample() so we're doing it by hand
order <- permute( 1:nObs )
trainSet <- passFailData[order[1:floor(nObs*.6)],1:7]
testSet <- passFailData[order[(floor(nObs*.6)+1):nObs],1:7]

model <- glm( formula = V1 ~ ., family=binomial, data=trainSet )

predLogit <- predict( model , newdata=testSet[,-1],
                     type="response", se.fit = FALSE )

#Lets do a confustion matrix based on a tau of .5
predictedValues <- floor( predLogit + .5 )

TP <- 0
FP <- 0
FN <- 0
TN <- 0

for( i in 1:(nObs-floor(.6*nObs)) ){
    if( (predictedValues[i] == (testSet$V1)[i]) &&
        (predictedValues[i] == 1) ){
        TP <- TP + 1
    }
    else if( predictedValues[i] != testSet$V1[i] &&
             predictedValues[i] == 1 ){
        FP <- FP + 1
    }
    else if( predictedValues[i] == testSet$V1[i] ){
        TN <- TN + 1
    }
    else {
        FN <- FN + 1
    }
}
TP

## [1] 653
```

```r
FP
```

```
## [1] 396
```

```r
FN
```

```
## [1] 1022
```

```r
TN
```

```
## [1] 1929
```

From this data we can build a confusion matrix.

653	396
1022	1929

```r
pred <- prediction(predLogit, testSet$V1)

auc <- performance(pred,"auc")
auc@y.values
```

```
## [[1]]
## [1] 0.6654476
```

```r
rocrCurve <- performance( pred,"tpr","fpr" )

plot( rocrCurve, xaxis="tpr", yaxis="fpr",
     main="Logistic Regression",
     sub=paste("AUC: ", toString(auc@y.values[1]) ) )
```

Logistic Regression