Alexa Summers

Gireesh Kumar Muppalla

Santhoshini Sree Bolisetty

Homework #3—CS 5565

1. $$\left.\begin{aligned} &Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y) \\ &Var(cX) = c^2 Var(X) \\ &Cov(cX,Y) = Cov(X,cY) = cCov(X,Y) \end{aligned}\right\} \text{Rules to use}$$

$$Var(\alpha X + (1-\alpha)Y) = Var(\alpha X) + Var((1-\alpha)Y) + 2Cov(\alpha X, (1-\alpha)Y)$$

$$= \alpha^2 Var(X) + (1-\alpha)^2 Var(Y) + 2\alpha(1-\alpha) Cov(X,Y)$$

$$= \sigma^2_x \alpha^2 + \sigma^2_y (1-\alpha)^2 + 2\sigma_{xy}(-\alpha^2 + \alpha)$$

$$\frac{d}{d\alpha} f(\alpha) \Rightarrow$$

$$0 = 2\sigma^2_x \alpha + 2\sigma^2_y (1-\alpha)(-1) + 2\sigma_{xy}(-2\alpha + 1)$$

$$0 = \sigma^2_x \alpha + \sigma^2_y(\alpha - 1) + \sigma_{xy}(-2\alpha + 1)$$

$$0 = (\sigma^2_x + \sigma^2_y - 2\sigma_{xy})\alpha - \sigma^2_y + \sigma_{xy}$$

$$\boxed{\alpha = \frac{\sigma^2_y - \sigma_{xy}}{\sigma^2_x - \sigma^2_y - 2\sigma_{xy}}}$$

2a. The jth observation has a probability of 1/n of being the first bootstrap sample, so the probability that it is not the first bootstrap sample is 1-1/n.

2b. The jth observation has a probability of 1/n of being the second bootstrap sample, so the probability that it is not the second bootstrap sample is 1-1/n.

2c. $(1-1/n)(1-1/n)(1-1/n)\ldots = (1-1/n)^n$. Each observation has an independent chance of equaling the jth, so after applying the product rule, we end up with $(1-1/n)^n$.

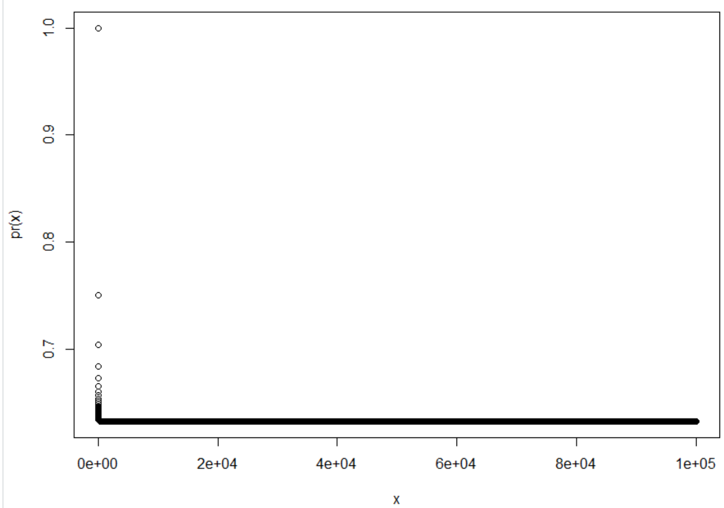
2d. n = 5 -> $1-(1-1/5)^5=1-(4/5)^5=67.2\%$

2e. n = 10 ->$1-(1-1/100)^{10}=1-(99/100)^{100}=63.4\%$

2f. n = 10,000 -> $1-(1-1/10000)^{10000}=63.2\%$

2g.

```
> pr = function(n) return(1 - (1 - 1/n)^n)
> x = 1:1e+05
> plot(x, pr(x))
```



2h.

```
> store=rep(NA, 10000)
> for(i in 1:10000) {
+ store[i]=sum(sample (1:100, rep=TRUE)==4) >0
+ }
> mean(store)
[1] 0.6329
```

The asymptote is 63.29% and is approached rapidly. The selection probability is almost equal to the answer for 2f (n = 10,000 -> $1-(1-1/10000)^{10000}$=63.2%), as expected.

3a. Take a set of observations and split them into non-overlapping groups (k). These groups will act as the remainder of a testing set. When the resulting MSE estimates are averaged together, the test error can be estimated.

3bi. Advantages: easy to implement and relatively simple to understand.

Disadvantages: The estimation of the test error can have a high variation depending on observations in the training and testing sets. Only a subset of observations are used to fit the model, which tends to result in a worse performance.

3bii. Advantages: Less bias, less variable MSE.

Disadvantages: Computationally intensive—takes a long time.

4. We could use a bootstrap distribution by sampling observations from the original dataset, and then fitting a new model for each repetition, and then looking at the RMSE of all the estimates.

5.

```r
library (MASS)
library (ISLR)

X = c(1,2,3,4,5,6,7,8,9,10)
Y = c(1.00, 2.00, 1.3, 3.75, 2.25, 4.5, 5.21, 4.98, 6.26, 5.4)

df = data.frame(X, Y)
print (df)

plot (Y ~ X, data = df)

plot(Y ~ X, data = df, col = c("green"))
cor(df)
fit = lm (Y ~ X, data = df)

#RSS (Sum of squared residuals)
deviance(fit)
sum(resid(fit)^2)

summary(fit)
```

```
> X = c(1,2,3,4,5,6,7,8,9,10)
> Y = c(1.00, 2.00, 1.3, 3.75, 2.25, 4.5, 5.21, 4.98, 6.26, 5.4)
> df = data.frame(X, Y)
> print (df)
    X    Y
1   1 1.00
2   2 2.00
3   3 1.30
4   4 3.75
5   5 2.25
6   6 4.50
7   7 5.21
8   8 4.98
9   9 6.26
10 10 5.40
> plot (Y ~ X, data = df)
> plot(Y ~ X, data = df, col = c("green"))
> cor(df)
          X         Y
X 1.0000000 0.9186152
Y 0.9186152 1.0000000
> fit = lm (Y ~ X, data = df)
> #RSS (Sum of squared residuals)
> deviance(fit)
[1] 5.002115
> sum(resid(fit)^2)
[1] 5.002115
> summary(fit)

Call:
lm(formula = Y ~ X, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.1288 -0.6597  0.1247  0.5808  0.9436 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.51667    0.54018   0.956 0.366838    
X            0.57242    0.08706   6.575 0.000174 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7907 on 8 degrees of freedom
Multiple R-squared:  0.8439,	Adjusted R-squared:  0.8243 
F-statistic: 43.23 on 1 and 8 DF,  p-value: 0.0001738
```

Coefficient of $\hat{B}_0$: 0.51667

Coefficient of $\hat{B}_1$: 0.57242

RSS: 5.002115

RSE: 0.7907

R^2 : 0.8439

T-statistic: 6.575

P-Value: .000174

Reject the null hypothesis.

6.

```r
X1 = c(1,2,3,4,5,6,7,8,9,10)
Y1 = c(1.21, 1.98, 4.76, 3.9, 6.2, 7.14, 9.35, 8.24, 10.16, 12.2)

df1 = data.frame(X1, Y1)
print (df1)

plot (Y1 ~ X1, data = df1)

plot(Y1 ~ X1, data = df1, col = c("green"))
cor(df1)
fit1 = lm (Y1 ~ X1, data = df1)

#RSS (Sum of squared residuals)
deviance(fit1)
sum(resid(fit1)^2)

summary(fit1)
```

```
> Y1 = c(1.21, 1.98, 4.76, 3.9, 6.2, 7.14, 9.35, 8.24, 10.16, 12.2)
> df1 = data.frame(X1, Y1)
> print (df1)
   X1    Y1
1   1  1.21
2   2  1.98
3   3  4.76
4   4  3.90
5   5  6.20
6   6  7.14
7   7  9.35
8   8  8.24
9   9 10.16
10 10 12.20
> plot (Y1 ~ X1, data = df1)
> plot(Y1 ~ X1, data = df1, col = c("green"))
> cor(df1)
          X1        Y1
X1 1.0000000 0.9766181
Y1 0.9766181 1.0000000
> fit1 = lm (Y1 ~ X1, data = df1)
> #RSS (Sum of squared residuals)
> deviance(fit1)
[1] 5.348956
> sum(resid(fit1)^2)
[1] 5.348956
> summary(fit1)

Call:
lm(formula = Y1 ~ X1, data = df1)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.16582 -0.46473 -0.02555  0.42664  1.13782 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.15200    0.55859   0.272    0.792    
X1           1.15673    0.09002  12.849 1.27e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8177 on 8 degrees of freedom
Multiple R-squared:  0.9538,	Adjusted R-squared:  0.948 
F-statistic: 165.1 on 1 and 8 DF,  p-value: 1.271e-06
```

Coefficient of $\hat{B}_0$: 0.15200

Coefficient of $\hat{B}_1$: 1.15673

RSS: 5.348956

RSE: .8177

R^2 : 0.9538

T-statistic: 12.849

P-Value: 1.27e-06

Reject the null hypothesis.