Homework 1

Gitanshu Munjal

October 9, 2014

1 Question 1

Write the formula for the complete simple linear regression model. Then, write a formula for a more complex model within which the simple linear regression is embedded. Use the typical symbols for response, predictor, and random variables and parameters. [20]

1.1 Complete Simple Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Where i identifies each of the n observations,
β_0 and β_1 are the intercept and slope respectively (parameters),
X_i is a constant of known value for the ith observation,
ϵ_i is a random variable with mean 0, variance σ^2, and uncorrelated with any other ϵ_i.

1.2 Complex Model within which the simple linear regression is embedded

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X^2 + \epsilon$$

Where symbols have their usual meaning and ϵ_i is Normal (0,σ^2)

---

2 Question 2

Using the file RegSim.R: Keep beta0 = 200 and beta1 = 50. Perform ¿100 simulations using the ”for loop” (line 75). Then for each iteration of the simulation make predictions for the yield expected for X=9 and X=18. Your output should be a matrix with two columns; one for predicted yield when X=9 and one for predicted yield when X=18. This matrix will have ¿100 rows. This should be automated with R. Display the head and tail of the matrix you created, and a histogram for each column. [30]

2.1 Code and Relevant Output

```
beta0<-200
beta1<-50
sigma<-300
x<-c(1,1,1,3,3,3,4,4,7,7,7,7,9,9,12,12,13,13,14,14,14,18,18,18)

bcoef<-matrix(0,1000,2)

for(i in 1:1000){
  newy      <-beta0+beta1*x+rnorm(length(x),0, sigma)
  new.slr   <-lm(newy~x)
  bcoef[i,]<-coef(new.slr)
}

yhat9            <-bcoef[,1]+bcoef[,2]*9
yhat18           <-bcoef[,1]+bcoef[,2]*18
yhatpredictions<-cbind(yhat9,yhat18)
```

```
head(yhatpredictions)
```

yhat9	yhat18
587.6195	969.6267
672.0038	1227.7465
722.9869	1088.0439
665.9900	1121.6247
747.1543	1297.7840
677.7975	1119.4805

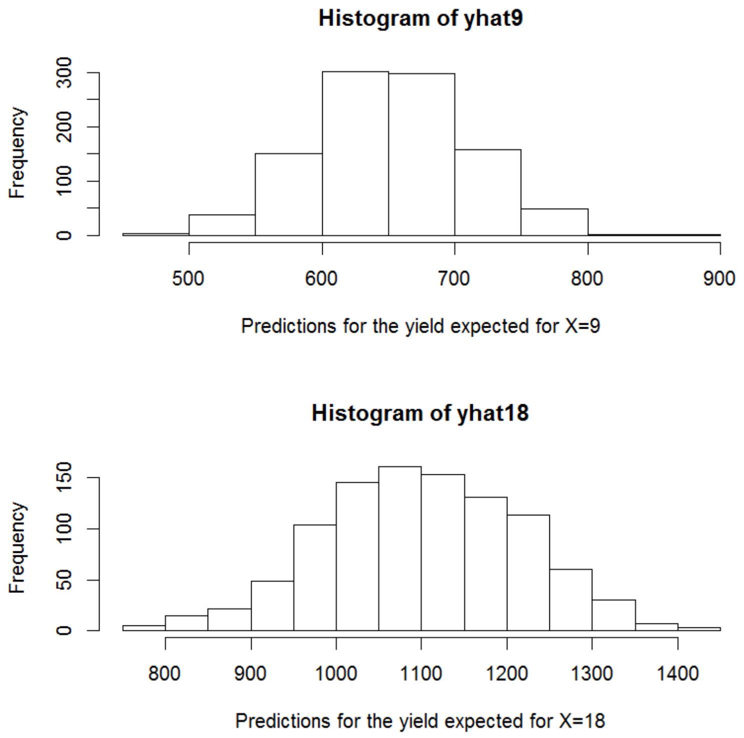
Table 1: Output from head(yhatpredictions) rows 1:6

```
tail(yhatpredictions)
```

yhat9	yhat18
691.2795	1211.303
523.6268	1048.479
654.6671	1197.683
560.1245	1000.698
678.0552	1096.819
619.5419	1190.620

Table 2: Output from tail(yhatpredictions) rows 995:1000

```
par(mfrow = c(2,1))
hist(yhat9, xlab=”Predictions␣for␣the␣yield␣expected␣for␣X=9”)
hist(yhat18, xlab=”Predictions␣for␣the␣yield␣expected␣for␣X=18”)
```



3 Question 3

Calculate the variance of each of the columns above using the equation to estimate the variance of a random variable. Check if the variance of each column obtained from the repeated experiments is consistent with the equations in the notes Ch2 and Faraway-PRA.pdf page 39. What relationship between the two methods would you have expected and why? (i.e. - Are they the same? One larger? Why?) [20]

3.1 Code and Relevant Output

3.1.1 Variance of each of the columns using the equation to estimate the variance of a random variable

```
((sum((yhat9-mean(yhat9))^2))/999)                  #hand calculation#
```

3566.504

```
var(yhat9)                                          #cross-check with R function
```

3566.504

```
((sum((yhat18-mean(yhat18))^2))/999)                #hand calculation#
```

13700.52

```
var(yhat18)                                         #cross-check with R function
```

13700.52

3.1.2 Variance of each of the columns using the equation to estimate the variance of prediction

```
MSE <-sigma^2
n   <-24
xh  <-cbind(9,18)
xbar<-mean(x)
a   <-(xh-xbar)^2
b   <-sum((x-xbar)^2)

est.var<-MSE *(1/n + (a/b))
colnames(est.var)<-c("X=9","X=18")
rownames(est.var)<-"Estimated Variance"
est.var<-as.matrix(est.var)
est.var
```

Output

```
                        X=9      X=18
Estimated Variance  3753.409  14062.5

comparison<-matrix()
comparison<-rbind(est.var,c(var(yhat9),var(yhat18)))
rownames(comparison)<-c("Est. variance as prediction","Est. variance as random variable")
comparison
```

Output

```
                                      X=9      X=18
Est. variance as prediction        3753.409  14062.50
Est. variance as random variable   3566.504  13700.52
```

The comparison makes it clear that the estimated variance as a random variable for a given level of X is lower than the estimated variance as prediction. This is consistent with expectation as the equation for the latter has more uncertainty terms and hence more variable. It is also important to note that within the type of estimated variance the estimates are much more variable at higher levels of X(18) compared to lower levels(9) which is also consistent with the equation.

4 Question 4

Assume that each simulation represents a real experiment. Describe two ways to create a confidence interval for beta1. Note that each simulation run can yield estimates of beta1 and its CI. This question is not necessarily based on the greater than 100 simulations you obtained for question 3, but it applies for any set of simulations or experiments. Any two methods different from each other will receive credit if they result in reasonable estimates of the CI. [20]

4.1 Code and Relevant Output

4.1.1 Using quantiles

```
qbounds<-as.matrix(quantile(bcoef[,2],c(0.025,0.975)))
colnames(qbounds)<-"95percent␣confidence␣level"
rownames(qbounds)<-c("Lower␣Bound","Upper␣Bound")
qbounds
```

Output

```
            95percent confidence level
Lower Bound                   27.67706
Upper Bound                   70.98168
```

4.1.2 Using student-t based equation

```
UpperBound<-mean(bcoef[,2])+sd(bcoef[,2])*qt(0.975,22)
LowerBound<-mean(bcoef[,2])-sd(bcoef[,2])*qt(0.975,22)
tbounds<-as.matrix(c(LowerBound,UpperBound))
colnames(tbounds)<-"95percent␣confidence␣level"
rownames(tbounds)<-c("Lower␣Bound","Upper␣Bound")
tbounds
```

Output

```
            95percent confidence level
Lower Bound                   27.41588
Upper Bound                   71.06007
```

5 Question 5

Assume that there is only one sample given by the data in PfertParSimData.csv. Is there a significant effect of fertilization on yield? What are the F-value and the corresponding p-value? [10]

5.1 Code and Relevant Output

```
fertdata<-read.table("C:\\Users\\gitanshu\\Desktop\\pfertparsimdata.csv",header=T,sep=",")
slr.fert<-lm(fertdata[,2]~fertdata[,1])
test<-anova(slr.fert)
Result<-as.matrix(c(anova(slr.fert)$"Pr(>F)"[1],anova(slr.fert)$"F␣value"[1]))
colnames(Result)<-"anova(slr.fert)"
rownames(Result)<-c("p-value","F-value")
Result
```

Output

```
           anova(slr.fert)
p-value        0.001409734
F-value       13.324458582
```

The results clearly show that there were significant differences between yields in response to fertilization levels as the p-value for the experiment is highly significant(p¡0.01).