Gawron_Homework10

Nicholas Gawron

4/12/2021

Problems

Question 1

Regression Estimation (Part A)

Note that our b calculation is nothing more than:

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)S_x^2}$$

The above calculation is also done by fetching the second coefficient in the lm model.

- Note that S_X^2 is calculated by computing `sd(dogtimess$x)^2`
- Our variable a comes from the line: `a<- mean(dogtimes$y)-b*mean(dogtimes$x)`
- `MuX` is our given μ for the subsidiary variable of 17
- Our population is $400 = N$
- The line `MuY <- mean(dogtimes$y)+b*(MuX-mean(dogtimes$x))` is analogous to $\mu_y = \bar{y} + b(\mu_x - \bar{x})$
- Our Mean Square error was calcuated by `MSE<- (sum((dogtimes$y - (a+b*dogtimes$x))^2))/(n-2)`
- Our variance for our estimate of μ_y, or $var(\mu_y)$ is `VarMuY <- (1-n/N)*(MSE/n); VarMuY`
- Our bound on the error of estimation is simply $B_{\mu_y} = 2\sqrt{var(\mu_y)}$. This is stored as `BMuY<-2*sqrt(VarMuY)`

```
StdDog <- c(14.3, 15.7 ,17.8 ,17.5 ,13.2 ,18.8 ,17.6 ,14.3 ,14.9 ,17.9 ,19.2)
NewDog<- c(15.2, 16.1, 18.1, 17.6, 14.5, 19.4, 17.5, 14.1, 15.2, 18.1, 19.1)

dogtimes <- as_tibble(data.frame(x = StdDog, y= NewDog))
ggplot(data =dogtimes, aes(x =x, y =y))+geom_point(pch =13, col= "red")+
  labs(title= "Dog Training Times", xlab = "Standard", ylab ="New Dog Training")
```
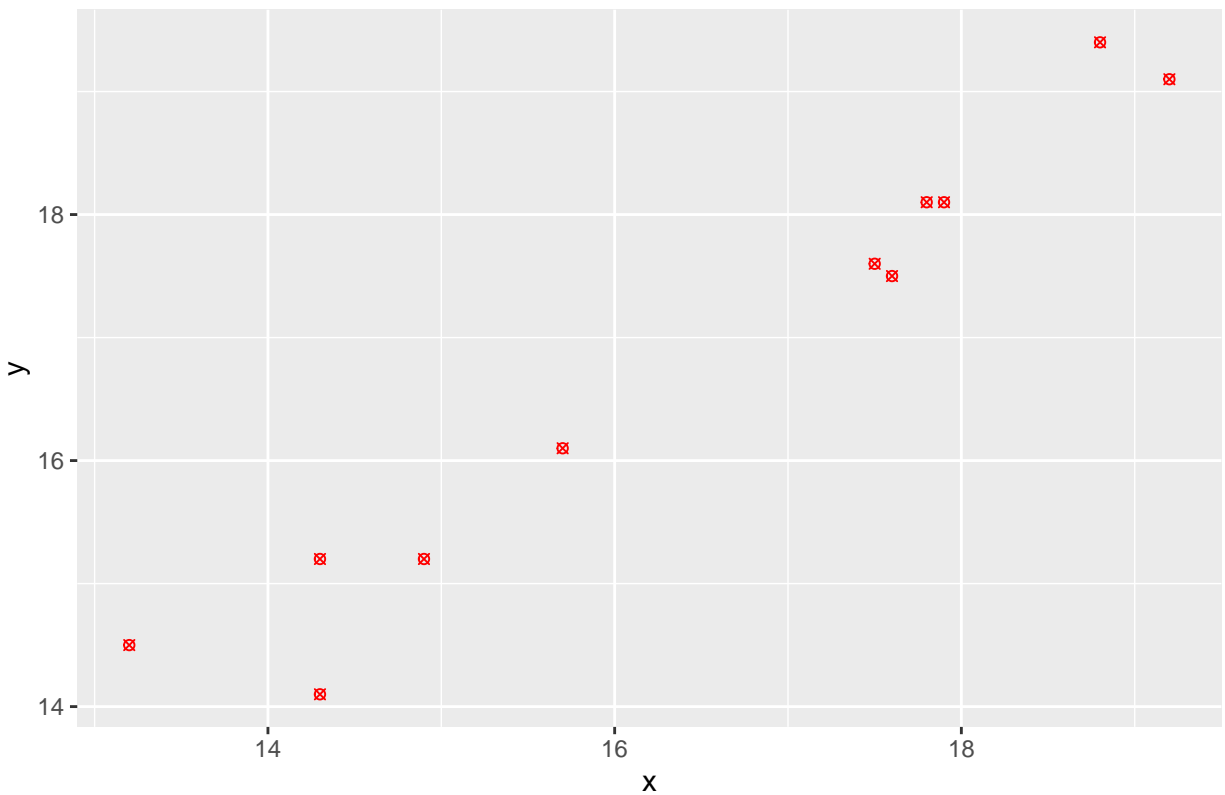
Dog Training Times

```r
MuX<-17 #given
N<- 700 #given population
n<-as.numeric(length((dogtimes$x))) #Sample size calculation

#calculates b
b <- sum((dogtimes$y - mean(dogtimes$y))*(dogtimes$x - mean(dogtimes$x)))/((n-1)*(sd(dogtimes$x)^2));b
```

[1] 0.8872302

```r
MuY <- mean(dogtimes$y)+b*(MuX-mean(dogtimes$x));MuY
```

[1] 17.2769

```r
a<- mean(dogtimes$y)-b*mean(dogtimes$x)

MSE<- (sum((dogtimes$y - (a+b*dogtimes$x))^2))/(n-2)
VarMuY <- (1-n/N)*(MSE/n); VarMuY
```

[1] 0.01519949

```r
BMuY<- 2*sqrt(VarMuY); BMuY
```

[1] 0.2465724

- Our output for μ_y is 17.2769032
- Our output for Var_y is 0.0151995
- Our output for B_{μ_y} is 0.2465724

```
model1<-lm(NewDog~StdDog)
a <-  model1$coefficients[1];a<-as.numeric(a);a
```

[1] 2.19399

```
b<-model1$coefficients[2];b<-as.numeric(b);b
```

[1] 0.8872302

Differnece Estimation (Part B)

We compute as followed:

- d of the difference vector is calculated by `d<- dogtimes$y-dogtimes$x`
- Our $\bar{d}$ is calculated by `dBAR<- mean(dogtimes$y)-mean(dogtimes$x)`
- Our mean estimate is: $\mu_y = \mu_x + \bar{d}$ stored as: `MuYDiff <- MuX+ dBAR;`
- Our variance is calculated by: `(1-n/N)*(1/n)*((sum((d-dBAR)^2)))/(n-1)`, this represents the formula: $(1 - \frac{n}{N})(\frac{1}{n})(\frac{\sum_i^n (d_i - \bar{d})^2}{n-1}) = Var(\mu_y)$

```
#Computes the difference variable d
d<- dogtimes$y-dogtimes$x;
dBAR<- mean(dogtimes$y)-mean(dogtimes$x);

MuYDiff <- MuX+ dBAR; MuYDiff
```

[1] 17.33636

```
VarMuYDiff <- (1-n/N)*(1/n)*((sum((d-dBAR)^2)))/(n-1)
VarMuYDiff
```

[1] 0.01848179

```
BMuYDiff <- 2*sqrt(VarMuYDiff); BMuYDiff
```

[1] 0.2718955

- Our output for $\mu_{y,D}$ is 17.3363636
- Our output for $Var_{y,D}$ is 0.0184818
- Our output for $B_{\mu_{y,D}}$ is 0.2718955

Evaluation (Part C)

We note that our bound for the regression estimation, which is 0.2465724, is less than our bound on the error of estimation for difference calculations. Our bound for the difference calculations is: 0.2718955. Since the bound for regression estimation is smaller than that of ratio estimation - we would prefer to use regression estimation methods. This is because we are no longer using the slope on our regression coefficient, since in difference estimation. This is because we set the 'slope' in difference estimation to a value of one. This is problematic because our estimated slop is about 2.19 which differs greatly from 1. Despite difference estimation being easier - it gave us a worse bound and would be less prefered.

Question 2

Regression

Total In Regression:

$$\mu_{Y,L} = \bar{y} + b(\mu_x - \bar{x})$$
$$N\mu_{Y,L} = N(\bar{y} + b(\mu_x - \bar{x}))$$
$$\tau_{Y,L} = N\bar{y} + b(\tau_x - N\bar{x})$$

Variance in model:

$$Var(\mu_{Y,L}) = (1 - \frac{n}{N})(\frac{MSE}{n})$$
$$(N^2)Var(\mu_{Y,L}) = (N^2)(1 - \frac{n}{N})(\frac{MSE}{n})$$
$$Var(N\mu_{Y,L}) =$$
$$Var(\tau_{Y,L}) = (N^2)(1 - \frac{n}{N})(\frac{MSE}{n})$$

Difference estiamation

Total:

$$\mu_{y,D} = \bar{y} + (\mu_x - \bar{x}) = \mu_x + \bar{d}$$
$$N\mu_{y,D} = N(\bar{y} + (\mu_x - \bar{x}) = \mu_x + \bar{d})$$
$$\tau_{y,D} = N\bar{y} + (\tau_x - N\bar{x}) = \tau_x + N\bar{d}$$

Variance:

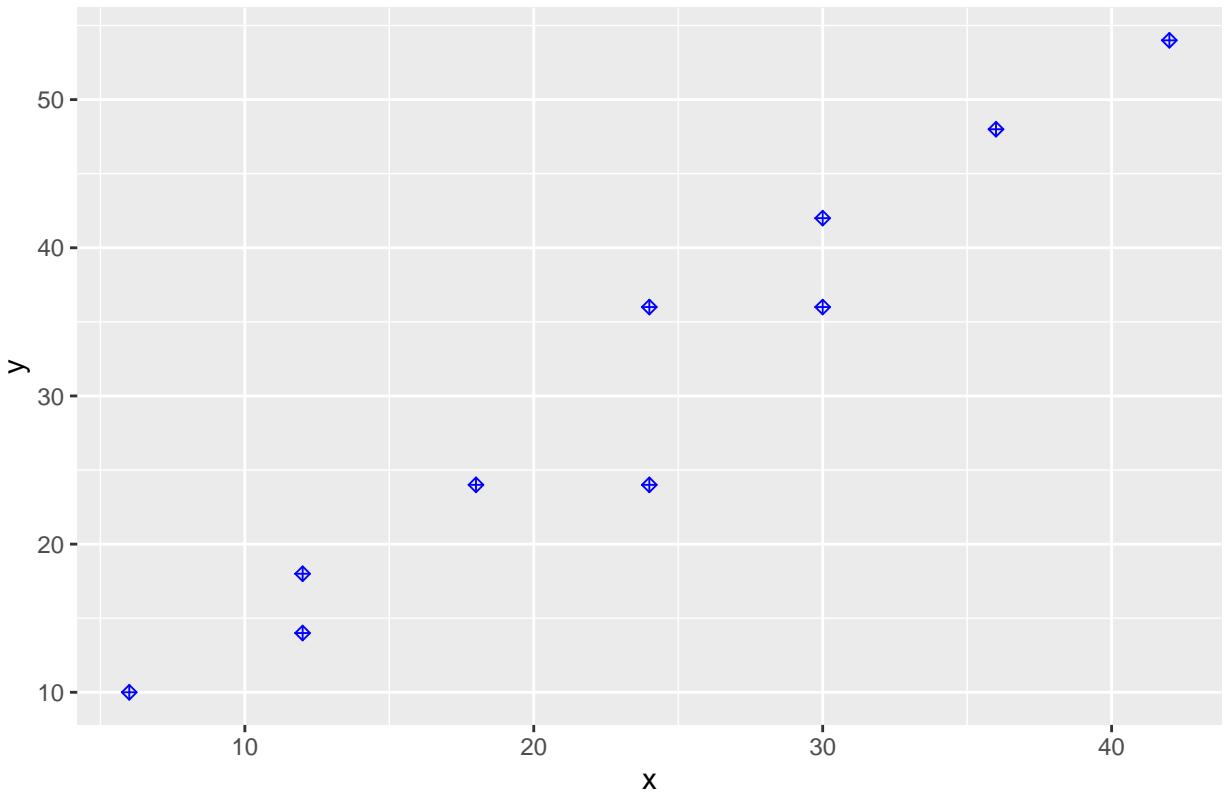$$Var(\tau_{Y,D}) = Var(N\mu_{Y,D})$$
$$= N^2(Var(\mu_{Y,D}))$$
$$= N^2(1 - \frac{n}{N})(\frac{1}{n})(\frac{\sum_i^n (d_i - \bar{d})^2}{n - 1})$$

Question 3

Dead Firs Regression (Part A)

```
Pct <- c(12, 30, 24, 24, 18, 30, 12, 6, 36, 42)
Act <- c(18 ,42 ,24 ,36 ,24 ,36, 14 ,10 ,48 ,54)
DeadFirs <- as_tibble(data.frame(x = Pct, y= Act))
ggplot(data =DeadFirs, aes(x =x, y =y))+geom_point(pch =9, col= "blue")+
  labs(title= "DeadFirs Photo Vs. Actual", xlab = "Photo", ylab ="Actual")
```

DeadFirs Photo Vs. Actual


```r
model1<-lm(Act~Pct)
# This creates
a <-  model1$coefficients[1];a<-as.numeric(a);a
```

```
## [1] 1.130699
```

```r
b<-model1$coefficients[2];b<-as.numeric(b);b
```

```
## [1] 1.259372
```

```r
MuX<-4200/200 #given Mu
N<- 200 #given population  of acres
n<-as.numeric(length((DeadFirs$x))) #Sample size calculation

#calculates b
b <- sum((DeadFirs$y - mean(DeadFirs$y))*(DeadFirs$x - mean(DeadFirs$x)))/((n-1)*(sd(DeadFirs$x)^2))


MuY <- mean(DeadFirs$y)+b*(MuX-mean(DeadFirs$x));
TauY <- N*MuY; TauY
```

```
## [1] 5515.502
```

```r
a<- mean(DeadFirs$y)-b*mean(DeadFirs$x)

MSE<- (sum((DeadFirs$y - (a+b*DeadFirs$x))^2))/(n-2)
VarMuY <- (1-n/N)*(MSE/n); VarMuY
```

```
## [1] 1.257812
```

```
VarTot<- (N^2)*VarMuY
BTotY<- 2*sqrt(VarTot); BTotY
```

```
## [1] 448.6088
```

- Our estimated total is: 5515.5015198
- Our variance for the total is: 5.0312462×10^4
- Our bound on the error of estimation for the total is: 448.6087917
- It is important to note our regression values: a = 1.1306991 and b = 1.2593718

Conclusion on Question 3 (Part B)

We note that our bound on the error of estimation that was calculated under ratio estimation is `428.4381371`. We will assume this value is correct. We further note that our bound on the error of estimation for ratio estimation is less than our bound created in the previous part (448.6087917). Due to the fact that the bound of error is less for ratio estimation total than it was for regression estimation of the total - we conclude that ratio estimation is more appropriate here.

Question 4

It is important to note that we have do **not** have a randomly ordered data set. In particular this mortgage data is ordered with a general positive trend across serial number - since prices rose year by year.

A systematic sample for a random population behaves, for all practical purposes, like a simple random sample. So, in that case, the variance approximation using the formula from simple random sampling works well. In samples from an ordered population, the sample values will tend to be further apart numerically than in a simple random sample, making the within-sample correlation, ρ, negative. If we took a systematic sample for the mortgage data; each sample will have some of the smaller values as well as some of the larger values, which would not necessarily happen in a simple random sample. This can be because a simple random sample - while unlikely - could only take on values from the lower end of mortgage data and not give an accurate representation of the population. This implies that the systematic sampling mean will have a smaller variance than the one for simple random sampling, so that the use of the simple random sampling formula produces an overestimate of the true sampling error. Since systematic will gather data more uniformly and yeild a smaller variance - we will rather imploy systematic sampling.