Lab4

Eunjin Park

date

DELETE ANYTHING FROM THIS TEMPLATE BELOW THAT IS NOT PART OF YOUR SOLUTION.

(0) Instructions for installing tinytex for PDF rendering: **https://yihui.org/tinytex/**

```
install.packages('tinytex')
tinytex::install_tinytex()
```

I. First Model

(1) It's usually a good idea to change any categorical explanatory variables to factors before implementing linear regression in R. Provide executable code (PEC) that changes the MPA variable to a factor. Which of the 5 areas is assigned to the 3rd level?

```
Lob = read.csv("6dd93320.csv")

Lob$MPA = as.factor(Lob$MPA)

levels(Lob$MPA)[3]
```

```
## [1] "Point Vicente State Marine Conservation Area"
```

(2)Which variable gives the density of giant kelp? Proportion of 'feather boa' kelp? Of all the species of kelp in the dataset, which do you like the best? Why? -> Marco_dens

(3)For each of the following explanatory variables, give the variables type (continuous or categorical) and number of levels if categorical: "MPA", "Inside_Outside", "Depth_m", "Relief_cm", "Flat_Rock", "Cobble", "Boulder", "Sand"

```
#str(Lob)
#table(Lob$MPA)
#variable1 = as.factor(Lob$Inside_Outside)
#class(variable1)
#table(variable1)
#levels(variable1) #how many levels
```

```
set = unique(Lob$MPA)
length(set)
```

```
## [1] 5
```

```r
set = unique(Lob$Cobble)
length(set)
```

```
## [1] 72
```

```r
class(Lob$Inside_Outside)
```

```
## [1] "character"
```

```r
class(Lob$Depth_m)
```

```
## [1] "numeric"
```

```r
class(Lob$Relief_cm)
```

```
## [1] "numeric"
```

```r
class(Lob$Flat_Rock)
```

```
## [1] "numeric"
```

```r
class(Lob$Boulder)
```

```
## [1] "numeric"
```

```r
class(Lob$Sand)
```

```
## [1] "numeric"
```

-> Regarding MPA and Cobble both variables are categorical. And rest of the variables are continuous.

(4) Use R to create the model matrix, X, for a multiple regression model that uses the variables in (3) as predictors (PEC). What are the dimensions of X? Explain in your own words why the number of columns is not equal to the number of variable names, 8.

```r
Lob_form = Lob_total ~ MPA + Inside_Outside + Depth_m + Relief_cm + Flat_Rock + Cobble + Boulder + Sand

X = model.matrix(Lob_form, data = Lob, contrasts.arg = NULL, xlev = NULL)
X[1,]
```

```
##                                       (Intercept)
##                                              1.00
##            MPALaguna Beach State Marine Reserve
##                                              0.00
## MPAPoint Vicente State Marine Conservation Area
##                                              0.00
##          MPASouth La Jolla State Marine Reserve
```

```
##                                          0.00
##       MPASwami's State Marine Conservation Area
##                                          0.00
##                            Inside_OutsideOutside
##                                          0.00
##                                      Depth_m
##                                          6.80
##                                     Relief_cm
##                                         14.60
##                                     Flat_Rock
##                                         61.75
##                                       Cobble
##                                          1.50
##                                     Boulder
##                                       16.75
##                                        Sand
##                                       15.00
```

```
dim(X)
```

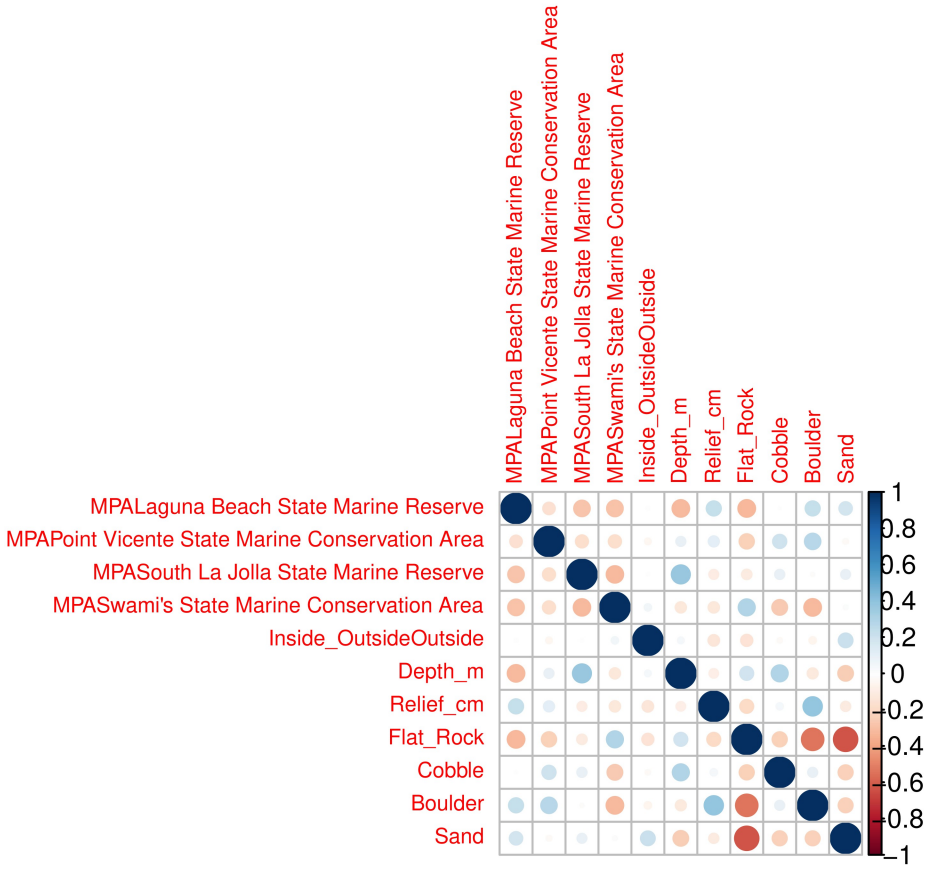
```
## [1] 162  12
```

(5)PEC that produces the correlation plot shown below (hint: you may want to adjust the tl.cex argument in corrplot()). Which two variables show the strongest correlation? What is their correlation coefficient?

```
#install.packages("corrplot")
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
Lob_data = Lob[ , c(1:9)]
corrplot(cor(X[ , -1]), tl.cex = 0.7)
```



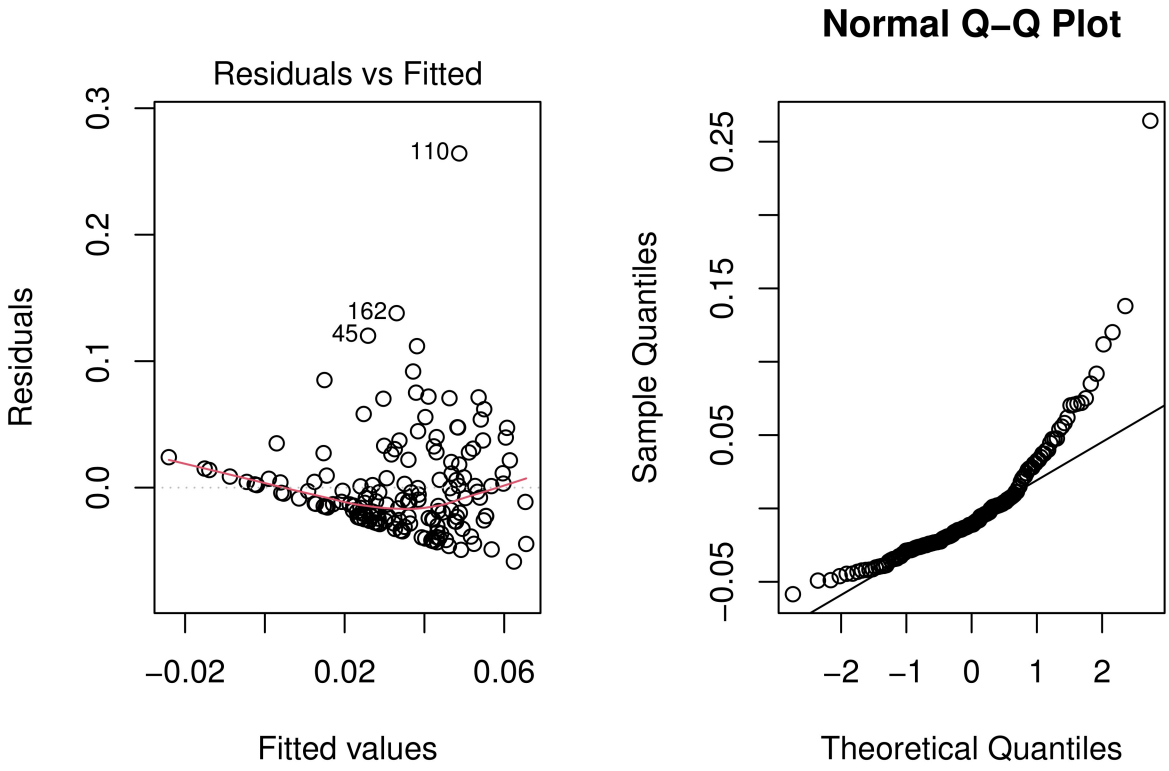

```
cor(Lob$Sand, Lob$Flat_Rock)
```

```
## [1] -0.6288679
```

-> Sand and Flat_Rock have strongly negative correlation.

(6) PEC that fits a linear regression model for the density of lobsters as a function of the explanatory variables in (3). Create two different plots that help check the assumption that the residuals are identically distributed and normal. What do your plots suggest? -> According to qqplot, there are outliers and each scatters are not normally distributed.

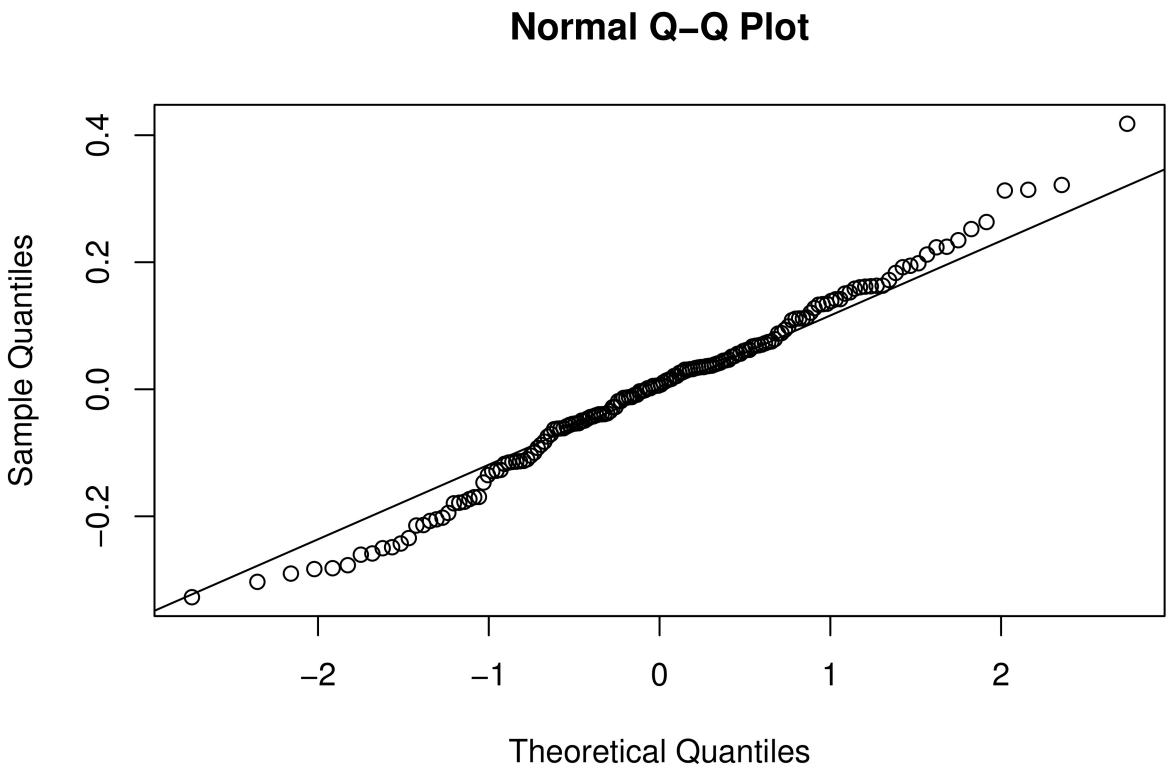
```
Lob_lm = lm(Lob_dens ~ MPA + Inside_Outside + Depth_m + Relief_cm + Flat_Rock + Cobble + Boulder, data =
layout(matrix(2:1, ncol = 2))
qqnorm(resid(Lob_lm))
qqline(resid(Lob_lm))
plot(Lob_lm, which = 1)
```



(7)Now fit a new model with the same predictors to the cube-root, y 1/3, of lobster density and re-create the two plots from (6). What do your new plots suggest?

-> Residuals for cube-root prediction is normally distributed.

```
Lob_cube = lm(Lob_dens^(1/3) ~ MPA + Inside_Outside + Depth_m + Relief_cm + Flat_Rock + Cobble + Boulde

qqnorm(resid(Lob_cube))
qqline(resid(Lob_cube))
```



(8)Which continuous explanatory variable has the largest estimated effect in magnitude on the cuberoot of lobster density? For which continuous explanatory variable do we have the greatest evidence that the effect is non-zero?

-> MPA point

```
summary(Lob_cube)
```

```
##
## Call:
## lm(formula = Lob_dens^(1/3) ~ MPA + Inside_Outside + Depth_m +
##     Relief_cm + Flat_Rock + Cobble + Boulder, data = Lob)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32731 -0.08047  0.00664  0.07799  0.41800
##
## Coefficients:
##                                                 Estimate Std. Error t value
## (Intercept)                                    2.827e-01  5.204e-02   5.432
## MPALaguna Beach State Marine Reserve          -9.001e-02  3.807e-02  -2.364
## MPAPoint Vicente State Marine Conservation Area -2.347e-01  4.957e-02  -4.736
## MPASouth La Jolla State Marine Reserve         3.536e-02  3.657e-02   0.967
## MPASwami's State Marine Conservation Area      -2.361e-02  3.405e-02  -0.693
## Inside_OutsideOutside                          1.444e-02  2.367e-02   0.610
## Depth_m                                       -1.537e-02  4.149e-03  -3.704
## Relief_cm                                     -2.943e-05  2.913e-04  -0.101
```

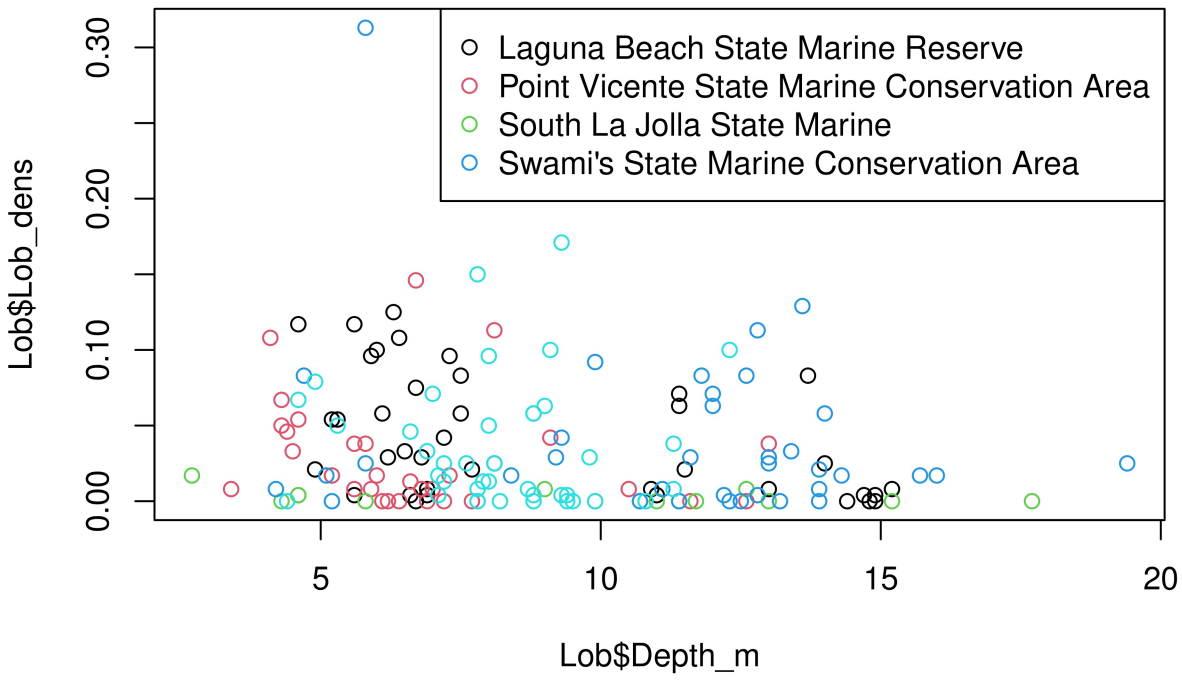
```
## Flat_Rock                                          1.387e-03  4.907e-04   2.826
## Cobble                                             7.956e-04  1.073e-03   0.741
## Boulder                                            3.057e-03  6.601e-04   4.631
##                                                   Pr(>|t|)
## (Intercept)                                       2.18e-07 ***
## MPALaguna Beach State Marine Reserve              0.019348 *
## MPAPoint Vicente State Marine Conservation Area 4.99e-06 ***
## MPASouth La Jolla State Marine Reserve            0.335122
## MPASwami's State Marine Conservation Area         0.489089
## Inside_OutsideOutside                             0.542732
## Depth_m                                           0.000297 ***
## Relief_cm                                         0.919674
## Flat_Rock                                         0.005357 **
## Cobble                                            0.459570
## Boulder                                           7.81e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1441 on 151 degrees of freedom
## Multiple R-squared:  0.3087, Adjusted R-squared:  0.263
## F-statistic: 6.744 on 10 and 151 DF,  p-value: 1.234e-08
```

(9)▪ Plot the density of lobsters as a function of depth, and color each point based on its associated MPA. Include a legend.3

```
plot(Lob$Depth_m, Lob$Lob_dens, col=c(1:5)[Lob$MPA], legend = levels(Lob$MPA))
```

```
## Warning in plot.window(...): "legend" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "legend" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "legend" is not a
## graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "legend" is not a
## graphical parameter

## Warning in box(...): "legend" is not a graphical parameter

## Warning in title(...): "legend" is not a graphical parameter
```

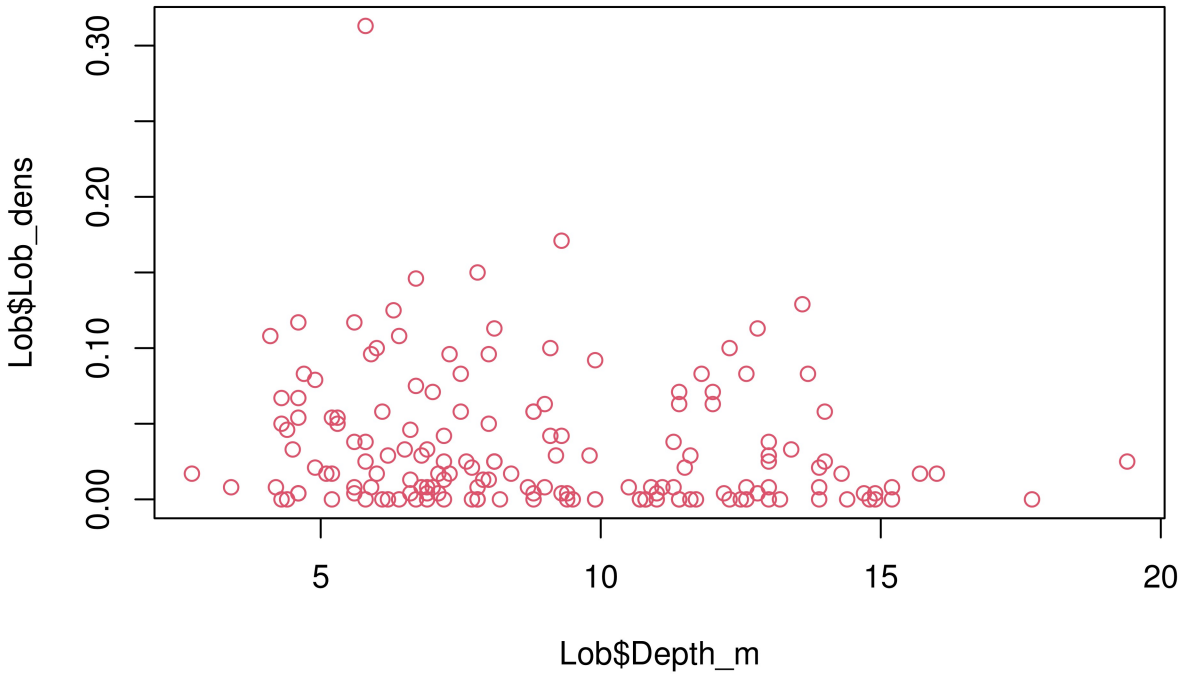
```
legend("topright" , legend = c("Laguna Beach State Marine Reserve","Point Vicente State Marine Conserva
```



▪ Add a curve that shows the predicted lobster density (not the cube-root of lobster density) for "Laguna Beach State Marine Reserve" inside the MPA, with all other variables set to their respective sample means. Color the curve to match your point color for "Laguna Beach State Marine Reserve" (PEC).

```
#Lob_lm = lm(Lob_dens ~ MPA + Inside_Outside + Depth_m + Relief_cm + Flat_Rock + Cobble + Boulder, data
#formula_log_transform <- log(stopduration) ~ exp_years + isschool + isguest +
#perceived_limited_english + perceived_age + perceived_gender + perceived_lgbt

#fit_log_transform= Lob_lm <- lm(formula_log_transform, data = ripa_2021)




#perceived_lgbt <- data.frame(exp_years = median(Lob$MPA),
#isschool = 1, isguest = 0,
#perceived_limited_english = 0,
#perceived_age = median(Lob$perceived_age),
#perceived_gender = "Female",
#perceived_lgbt = "Yes")

#exp(predict(fit_log_transform, newdata = perceived_lgbt, interval = "confidence"))

#exp_sequence <- 0:max(Lob$MPA)
#perceived_exp <- data.frame(Lob)
#predictions <- exp(predict(Lob_lm, newdata = perceived_exp,
#interval = "prediction"))
```

```
Lob_formula = Lob_dens ~ Depth_m
fit1 = lm(Lob_lm, data = Lob)
plot(Lob$Depth_m, Lob$Lob_dens, col=Lob$MPA[Lob$MPA == "Laguna Beach State Marine Reserve"])
```




```
CI = predict(Lob_lm, newdata = Lob, interval = "confidence")

#lines(x_axis, CI[ , 'upr'], col = 4, lwd = 2)
```

▪ Add another curve that shows the upper bound on pointwise 95% confidence intervals for the predicted density of lobster at a new site for each depth.

```
#plotCI = (x = Lob$Depth_m, y = Lob$Lob_dens , ui= Lob$upper)
```

II. Model section

(10) Begin with a model that includes all the variables used in section I. above, as well as the density of giant kelp and all percent-cover of the seaweed/algae variables: "MPA", "Inside_Outside", "Depth_m", "Relief_cm", "Flat_Rock", "Cobble", "Boulder", "Sand", "Macro_dens", "Pterygophora", "Laminaria", "Eisenia", "Egregia", "Cystoseira", "Plocamium", "Other_Red", "Coralline", "Surfgrass". Continue to use the cube-root of lobster density as a response. Use the step() function in R to select a model using backwards step selection based on AIC (PEC). Which variables are included?

```r
step(Lob_cube, scope = list(lower = ~ 1, upper = ~ MPA + Inside_Outside + Depth_m + Relief_cm + Flat_Ro
```

```
## Start:  AIC=-616.95
## Lob_dens^(1/3) ~ MPA + Inside_Outside + Depth_m + Relief_cm +
##     Flat_Rock + Cobble + Boulder
##
##                  Df Sum of Sq    RSS     AIC
## - Relief_cm       1   0.00021 3.1378 -618.94
## - Inside_Outside  1   0.00773 3.1453 -618.55
## - Cobble          1   0.01142 3.1490 -618.36
## <none>                        3.1376 -616.95
## - Flat_Rock       1   0.16590 3.3035 -610.60
## - Depth_m         1   0.28503 3.4226 -604.87
## - Boulder         1   0.44553 3.5831 -597.44
## - MPA             4   0.75735 3.8949 -589.92
##
## Step:  AIC=-618.94
## Lob_dens^(1/3) ~ MPA + Inside_Outside + Depth_m + Flat_Rock +
##     Cobble + Boulder
##
##                  Df Sum of Sq    RSS     AIC
## - Inside_Outside  1   0.00813 3.1459 -620.52
## - Cobble          1   0.01137 3.1492 -620.35
## <none>                        3.1378 -618.94
## - Flat_Rock       1   0.16570 3.3035 -612.60
## - Depth_m         1   0.28517 3.4230 -606.85
## - Boulder         1   0.48448 3.6223 -597.68
## - MPA             4   0.76856 3.9063 -591.45
##
## Step:  AIC=-620.52
## Lob_dens^(1/3) ~ MPA + Depth_m + Flat_Rock + Cobble + Boulder
##
##             Df Sum of Sq    RSS     AIC
## - Cobble     1   0.00980 3.1557 -622.02
## <none>                   3.1459 -620.52
## - Flat_Rock  1   0.15777 3.3037 -614.59
## - Depth_m    1   0.27762 3.4235 -608.82
## - Boulder    1   0.47635 3.6223 -599.68
## - MPA        4   0.77591 3.9218 -592.81
##
## Step:  AIC=-622.02
## Lob_dens^(1/3) ~ MPA + Depth_m + Flat_Rock + Boulder
##
##             Df Sum of Sq    RSS     AIC
## <none>                   3.1557 -622.02
## - Flat_Rock  1   0.14798 3.3037 -616.59
## - Depth_m    1   0.27316 3.4289 -610.57
## - Boulder    1   0.46823 3.6239 -601.61
## - MPA        4   0.76612 3.9218 -594.81

##
## Call:
## lm(formula = Lob_dens^(1/3) ~ MPA + Depth_m + Flat_Rock + Boulder,
```

```
##     data = Lob)
##
## Coefficients:
##                                       (Intercept)
##                                          0.295897
##            MPALaguna Beach State Marine Reserve
##                                         -0.093104
## MPAPoint Vicente State Marine Conservation Area
##                                         -0.235796
##          MPASouth La Jolla State Marine Reserve
##                                          0.030851
##       MPASwami's State Marine Conservation Area
##                                         -0.027279
##                                           Depth_m
##                                         -0.014146
##                                         Flat_Rock
##                                          0.001224
##                                           Boulder
##                                          0.002949
```

(11) Use the step() function again to select a model, but this time start with the model from section I. and take steps in both directions (PEC). Which variables were included in the best model using backward step selection, but are not included when taking steps in both directions?

```
step(Lob_lm, scope = list(lower= ~ 1, upper = ~MPA + Inside_Outside + Depth_m + Relief_cm + Flat_Rock +
```

```
## Start:  AIC=-1018.05
## Lob_dens ~ MPA + Inside_Outside + Depth_m + Relief_cm + Flat_Rock +
##     Cobble + Boulder
##
##                   Df Sum of Sq     RSS     AIC
## - Relief_cm        1 0.0000605 0.26389 -1020.0
## - Cobble           1 0.0004347 0.26426 -1019.8
## - Inside_Outside   1 0.0006491 0.26448 -1019.6
## - Flat_Rock        1 0.0026801 0.26651 -1018.4
## <none>                         0.26383 -1018.0
## - Boulder          1 0.0094747 0.27330 -1014.3
## - Depth_m          1 0.0122078 0.27604 -1012.7
## - MPA              4 0.0254527 0.28928 -1011.1
##
## Step:  AIC=-1020.01
## Lob_dens ~ MPA + Inside_Outside + Depth_m + Flat_Rock + Cobble +
##     Boulder
##
##                   Df Sum of Sq     RSS     AIC
## - Cobble           1 0.0004413 0.26433 -1021.7
## - Inside_Outside   1 0.0007036 0.26459 -1021.6
## - Flat_Rock        1 0.0026590 0.26655 -1020.4
## <none>                         0.26389 -1020.0
## + Relief_cm        1 0.0000605 0.26383 -1018.0
## - Boulder          1 0.0099396 0.27383 -1016.0
## - Depth_m          1 0.0122222 0.27611 -1014.7
## - MPA              4 0.0262024 0.29009 -1012.7
```

```
##
## Step:  AIC=-1021.74
## Lob_dens ~ MPA + Inside_Outside + Depth_m + Flat_Rock + Boulder
##
##                  Df Sum of Sq     RSS     AIC
## - Inside_Outside  1 0.0008126 0.26514 -1023.2
## <none>                        0.26433 -1021.7
## - Flat_Rock       1 0.0035245 0.26786 -1021.6
## + Cobble          1 0.0004413 0.26389 -1020.0
## + Relief_cm       1 0.0000671 0.26426 -1019.8
## - Boulder         1 0.0104352 0.27477 -1017.5
## - Depth_m         1 0.0154317 0.27976 -1014.5
## - MPA             4 0.0270566 0.29139 -1014.0
##
## Step:  AIC=-1023.24
## Lob_dens ~ MPA + Depth_m + Flat_Rock + Boulder
##
##                  Df Sum of Sq     RSS     AIC
## - Flat_Rock       1 0.0029256 0.26807 -1023.5
## <none>                        0.26514 -1023.2
## + Inside_Outside  1 0.0008126 0.26433 -1021.7
## + Cobble          1 0.0005503 0.26459 -1021.6
## + Relief_cm       1 0.0001303 0.26501 -1021.3
## - Boulder         1 0.0098791 0.27502 -1019.3
## - Depth_m         1 0.0148759 0.28002 -1016.4
## - MPA             4 0.0274585 0.29260 -1015.3
##
## Step:  AIC=-1023.47
## Lob_dens ~ MPA + Depth_m + Boulder
##
##                  Df Sum of Sq     RSS     AIC
## <none>                        0.26807 -1023.5
## + Flat_Rock       1 0.0029256 0.26514 -1023.2
## + Cobble          1 0.0013296 0.26674 -1022.3
## + Inside_Outside  1 0.0002137 0.26786 -1021.6
## + Relief_cm       1 0.0000721 0.26800 -1021.5
## - Boulder         1 0.0071775 0.27525 -1021.2
## - Depth_m         1 0.0127327 0.28080 -1018.0
## - MPA             4 0.0315976 0.29967 -1013.4


##
## Call:
## lm(formula = Lob_dens ~ MPA + Depth_m + Boulder, data = Lob)
##
## Coefficients:
##                                       (Intercept)
##                                         0.0636111
##            MPALaguna Beach State Marine Reserve
##                                        -0.0257408
## MPAPoint Vicente State Marine Conservation Area
##                                        -0.0446145
##          MPASouth La Jolla State Marine Reserve
##                                         0.0039345
##       MPASwami's State Marine Conservation Area
```

```
##                                    -0.0065133
##                                       Depth_m
##                                    -0.0029795
##                                       Boulder
##                                     0.0003335
```

(12)Make a plot of Cook's distances for each residual from the model fit in (11). With which MPA is the observation with the largest Cook's distance associated? What else is special/interesting about this observation? Do you see any reason why it should be removed from the analysis?

```
model = lm(Lob_dens ~ MPA + Inside_Outside + Depth_m + Relief_cm + Flat_Rock + Cobble + Boulder, data =
cooked_dist = cooks.distance(model)

layout(matrix(1:2, 1, 2))
plot(model , which = 4:5)
```