HW4: Mutiple Linear Regression (ALRM Ch. 6, 7)

MATH/STAT 571A

DUE: 10/20/2023 11:59pm

Homework Guidelines

Please submit your answers on Gradescope as a PDF with pages matched to question answers.

One way to prepare your solutions to this homework is with R Markdown, which provides a way to include mathematical notation, text, code, and figures in a single document. A template `.Rmd` file is available through D2L.

Make sure all solutions are clearly labeled, and please utilize the question pairing tool on Gradescope. You are encouraged to work together, but your solutions, code, plots, and wording should always be your own. Come and see me and/or our TA during office hours or schedule an appointment when you get stuck and can't get unstuck.

I. Mathematical Foundations [14 pts]

(1) [5 pts] (ALRM 6.22) For each of the following regression models, indicate whether it is a *general linear regression model*. If it is not, state whether it can be expressed in the form of (6.7) by a suitable transformation:

a. $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 \log_{10} X_{i2} + \beta_3 X_{i1}^2 + \varepsilon_i$ This is a general linear regression model.
b. $Y_i = \varepsilon_i \exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}^2)$ This is a nonlinear model that cannot be converted to the GLM form
c. $Y_i = \log_{10}(\beta_1 X_{i1}) + \beta_2 X_{i2} + \varepsilon_i$ Transform X_{i1} predictor into $10^{X_{i1}}$.
d. $Y_i = \beta_0 \exp(\beta_1 X_{i1}) + \varepsilon_i$ This is a nonlinear model that cannot be converted to the GLM form
e. $Y_i = \left[1 + \exp(\beta_0 + \beta_1 X_{i1} + \varepsilon_i)\right]^{-1}$ Transform the response Y as $log(Y_i^{-1} - 1) = \beta_0 + \beta_1 X_{i1} + \epsilon_i$

(2) [3 pts] (ALRM 6.25) An analyst wanted to fit the regression model $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i, i = 1, \ldots, n$, by the method of least squares when it is known that $\beta_2 = 4$. How can the analyst obtain the desired fit (i.e., make inference about β_0, β_1, β_3, and σ^2) by using a multiple regression computer program? Can maximize the likelihood of the unknown estimators β_0, β_1, and β_3 and input 4 for the β_2 for the MLE. The variance σ^2 can be estimated as MSE*(I-H) using the b estimators.

(3) [3 pts] (ALRM 7.35) Derive the relationship between β_k and β_k^* in (7.46a) for $p-1 = 2$. $$ \begin{aligned}

\end{aligned} $$

(4) [3 pts] Derive the relationship between β_k and β_k^* for both $k = 0$ and $k > 0$ for the alternative standardized model similar to (7.45) that does **not** involve standardizing the response or omitting the intercept (i.e., $Y_i = \beta_0^* + \beta_1^* X_{i1}^* + \cdots + \beta_{p-1}^* X_{i,p-1}^* + \varepsilon_i$ vs. $Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$)

II. Lobsters in Southern California [26 pts]

We will be using recently gathered data about the abundance of the California spiny lobster (*Panulirus interruptus*) along the southern coast of California. The California spiny lobster is an important commercial species for California, and there was concern that over-fishing might lead to a decline in the population of the species. In 2012, several marine protected areas (MPAs) were established where fishing is prohibited. The MPAs were not established explicitly to protect spiny lobsters, but they offer a natural experiment where abundance of spiny lobsters inside and outside the MPAs can be studied to see what effect the protected areas might have.

In 2012 and 2013, a team of researchers collected counts of spiny lobsters around 5 different MPAs along the coast near San Diego. Counts were made just inside and just outside each MPA. This **website** gives information about all the variables measured at each site, and this **website** gives the raw data (also posted on D2L). Your goal is to use linear models to study the relationships between the density of lobsters observed at each site and the characteristics of the sites.

(5) [2 pts] 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"`.

Continuous: "Depth_m", "Relief_cm", "Flat_Rock", "Cobble", "Boulder", "Sand"

"MPA" categorical: "Cabrillo State Marine Reserve", "Laguna Beach State Marine Reserve", "Point Vicente State Marine Conservation Area", "South La Jolla State Marine Reserve", "Swami's State Marine Conservation Area"

"Inside_Outside" categorical: "Inside", "Outside"

```
lobsters <- read.csv("./Datasets/lobsters_survey.csv")
lobsters_subset = lobsters[,c("MPA", "Inside_Outside", "Depth_m", "Relief_cm", "Flat_Rock", "Cobble", "
head(lobsters_subset)
```

```
##                              MPA Inside_Outside Depth_m Relief_cm Flat_Rock
## 1 Cabrillo State Marine Reserve         Inside     6.8      14.6     61.75
## 2 Cabrillo State Marine Reserve         Inside     6.1       3.0     77.25
## 3 Cabrillo State Marine Reserve         Inside     7.7      15.7     47.11
## 4 Cabrillo State Marine Reserve         Inside     6.5       2.0     65.50
## 5 Cabrillo State Marine Reserve         Inside     5.3       5.8     90.50
## 6 Cabrillo State Marine Reserve         Inside     5.6       4.0     57.00
##   Cobble Boulder  Sand
## 1   1.50   16.75 15.00
## 2   4.25    8.50 10.00
## 3   2.11   27.11 23.68
## 4   2.25   27.75  4.50
## 5   5.25    4.25  0.00
## 6   4.75    5.00 34.00
```

```
unique(c(lobsters$MPA, lobsters$Inside_Outside))
```

```
## [1] "Cabrillo State Marine Reserve"
## [2] "Laguna Beach State Marine Reserve"
## [3] "Point Vicente State Marine Conservation Area"
## [4] "South La Jolla State Marine Reserve"
## [5] "Swami's State Marine Conservation Area"
## [6] "Inside"
## [7] "Outside"
```

(6) [3 pts] Use R to create the model matrix, $\mathbf{X}$, for a multiple regression model that uses the variables in (5) as predictors. What are the dimensions of $\mathbf{X}$? Explain in your own words why the number of columns is not equal to the number of variable names, 8.

There are 12 columns: 1 for the intercept, 5 dummy categories for MPA, 2 for Inside_Outside, and 6 to account for each of the remaining continuous variables. This totals to 14. However, the baselines for each of the categorical variables are wrapped into the intercept, so you subtract 2. Then, the total is 14-2 = 12 which matches the dimensions of X here.

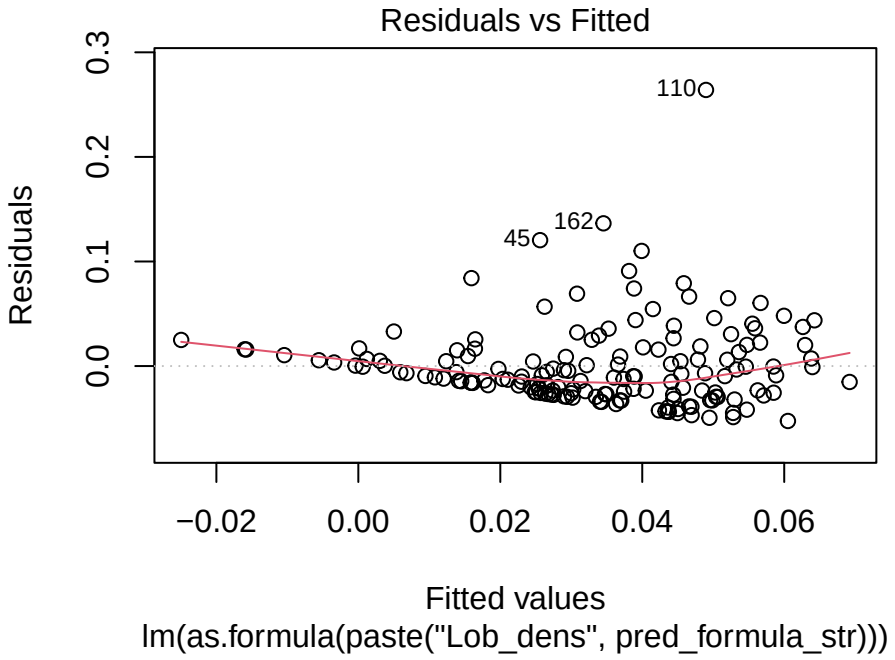
```
continuous <- names(lobsters_subset)[-c(1,2)]
pred_formula_str <- paste("~", paste(names(lobsters_subset), collapse='+'))
X <- model.matrix(as.formula(pred_formula_str), data = lobsters)
dim(X)
```

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

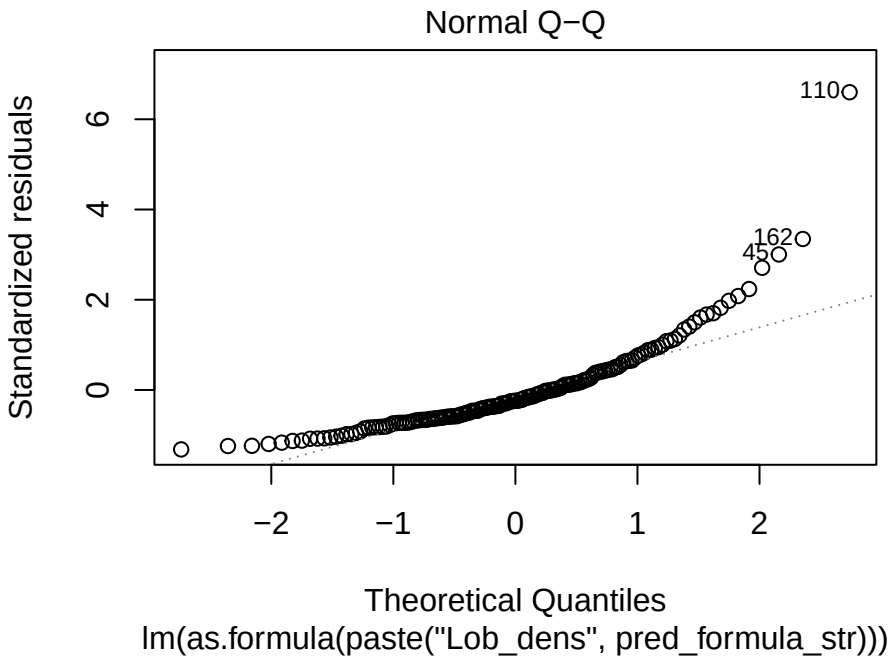
(7) [3 pts] Fit the linear regression model for the density of lobsters as a function of the explanatory variables in (5). Create two different plots that help check the assumptions that the residuals are identically distributed and normal. What do your plots suggest?

The residuals appear to be heteroskedastic and nonnormal. This means we should try a transform on the predicted variable to correct for both the residuals.

```
lob_dens_fit <- lm(as.formula(paste("Lob_dens", pred_formula_str)), data = lobsters)
plot(lob_dens_fit,1)
```



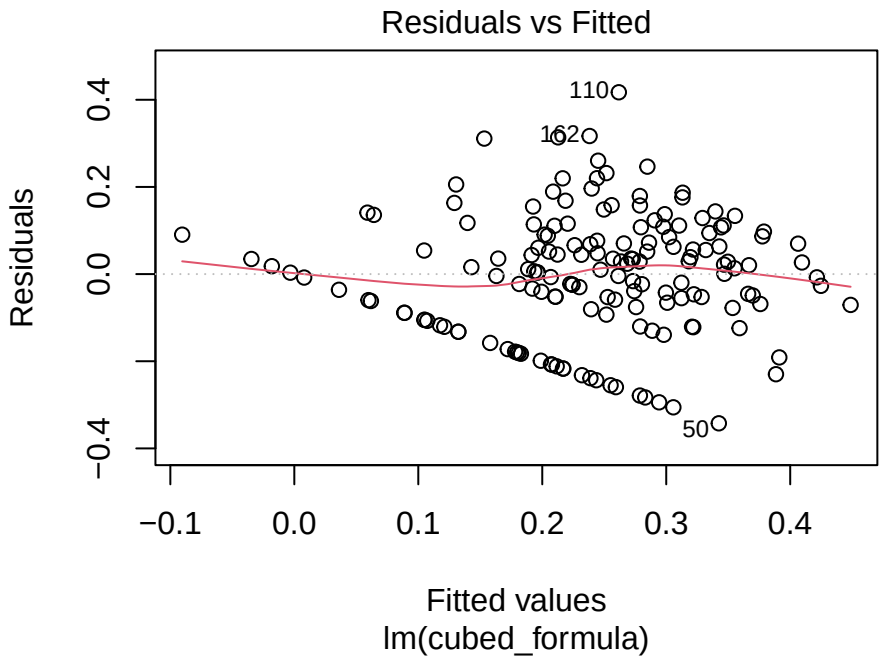
```
plot(lob_dens_fit,2)
```



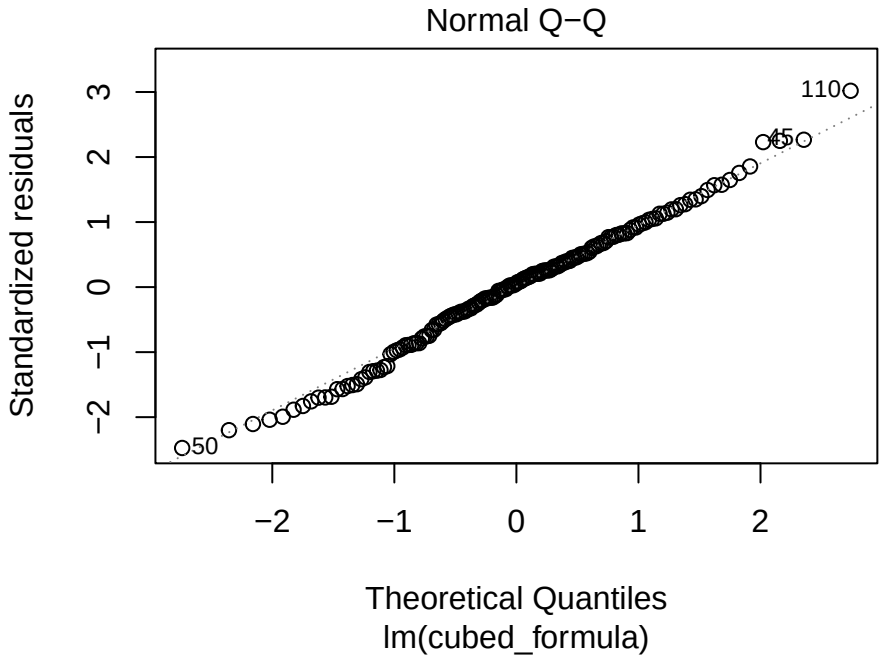
(8) [3 pts] Now fit a new model with the same predictors to the cube-root, $Y^{1/3}$, of lobster density and re-create the two plots from (7). What do your new plots suggest?

The new plot shows that the residuals are normally distributed and the residuals are less heteroskedastic.

```
cubed_formula <- as.formula(paste("Lob_dens^(1/3)", pred_formula_str))
lob_dens_rootfit <- lm(cubed_formula, data = lobsters)
plot(lob_dens_rootfit,1)
```




```
plot(lob_dens_rootfit,2)
```



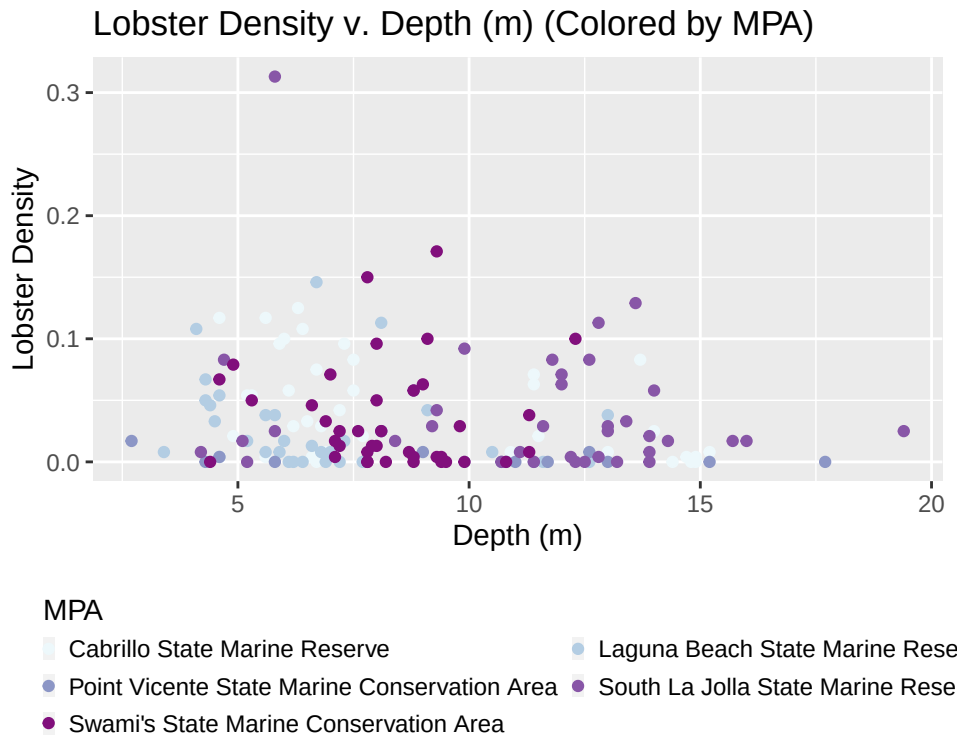
(9) Use the model from (8) to produce the following figure:

(i) [2 pts] Plot the density of lobsters as a function of depth and color each point based on its associated MPA. Include a legend.

```
library(ggplot2)
library(RColorBrewer)

color <- brewer.pal(n = 5, name = "BuPu")

legend_col <- scale_color_manual(values=c("Cabrillo State Marine Reserve" = color[1],
                            "Laguna Beach State Marine Reserve" = color[2],
                            "Point Vicente State Marine Conservation Area"=color[3],
                            "South La Jolla State Marine Reserve" = color[4],
                            "Swami's State Marine Conservation Area" = color[5]))
baseplot <- ggplot() +
  geom_point(aes(x = Depth_m, y = Lob_dens, color = MPA), data = lobsters) +
  legend_col +
  xlab("Depth (m)") +
  ylab("Lobster Density") +
  ggtitle("Lobster Density v. Depth (m) (Colored by MPA)") +
  theme(legend.key.size = unit(0.1, "cm"),
        legend.key.width = unit(0.1,"cm"),
        legend.position = 'bottom',
        legend.direction = 'vertical') +
  guides(color=guide_legend(nrow=3, byrow=TRUE))
baseplot
```

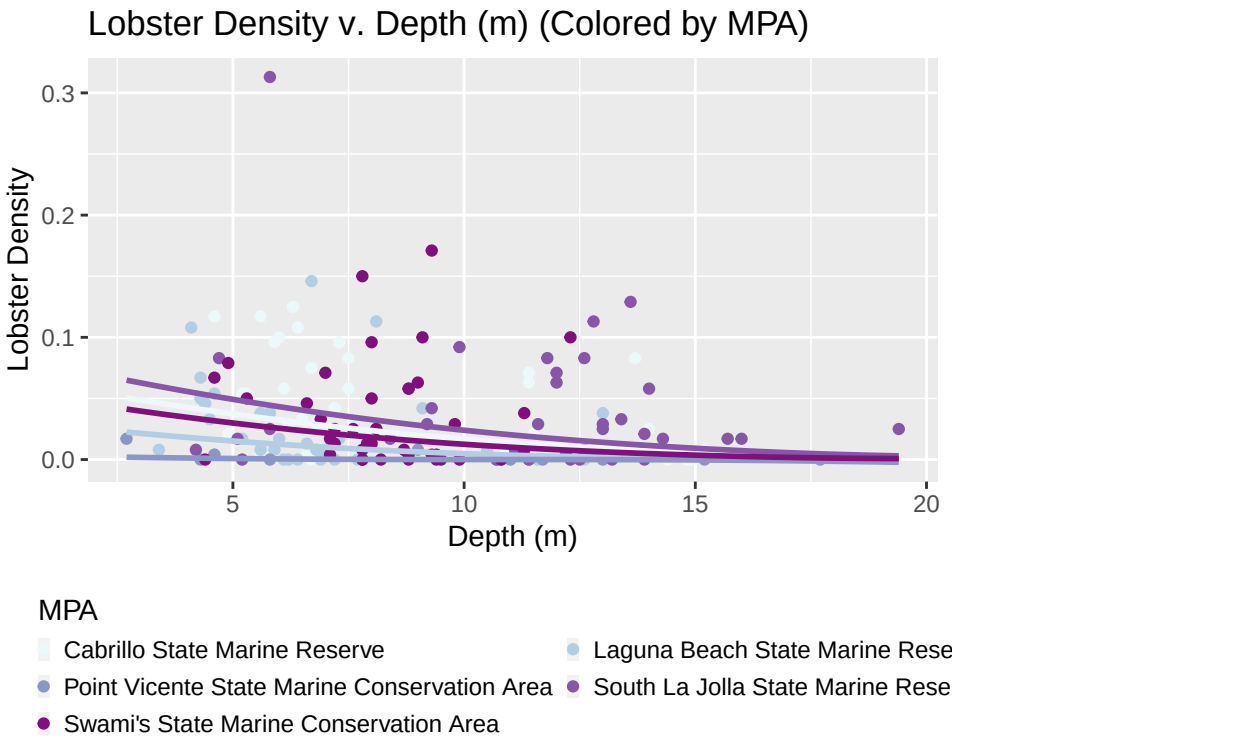


(ii) [2 pts] Add a curve that shows the predicted lobster density (***not*** *the cube-root of lobster density*) for "Laguna Beach State Marine Reserve" **inside** the MPA, with all other variables set to their respective sample means. Color the curve to match your point color for "Laguna Beach State Marine Reserve" (**PEC**).

```
mini <- min(lobsters$Depth_m)
maxi <- max(lobsters$Depth_m)
new_depthm <- seq(mini,maxi, .1)
n<-length(new_depthm)
new_data <- data.frame(Inside_Outside= rep("Inside",n),
                       Depth_m= new_depthm,
                       Relief_cm= rep(mean(lobsters$Relief_cm),n),
                       Flat_Rock= rep(mean(lobsters$Flat_Rock),n),
                       Cobble= rep(mean(lobsters$Cobble),n),
                       Boulder = rep(mean(lobsters$Boulder),n),
                       Sand = rep(mean(lobsters$Sand),n))

mpatrend <- function(mpa_type){
  trend <- predict(lob_dens_rootfit,
                   newdata = cbind(MPA = rep(mpa_type,n), new_data))
  trend_trans <- trend^3
  return(trend_trans)
}
mpa_type <- unique(lobsters$MPA)
plot2 <- baseplot +
  geom_smooth(aes(x=new_depthm, y = mpatrend(mpa_type[1])),
                color = color[1]) +
  geom_smooth(aes(x=new_depthm, y = mpatrend(mpa_type[2])),
                color = color[2]) +
  geom_smooth(aes(x=new_depthm, y = mpatrend(mpa_type[3])),
                color = color[3]) +
  geom_smooth(aes(x=new_depthm, y = mpatrend(mpa_type[4])),
                color = color[4]) +
  geom_smooth(aes(x=new_depthm, y = mpatrend(mpa_type[5])),
                color = color[5])
plot2
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



(iii) [4 pts] Add another curve that shows the upper bound on the Working-Hotelling 95% confidence band for the expected density of lobsters at each depth. Could not get this one to work, but here is my code attempt.

```
pred_WH_upper <- function(object, newdata, level = 0.95){
  fit <- predict(object, newdata) ## Yhat
  MSE <- summary(object)$sigma^2
  n <- nrow(object$model) ## extract df from model object
  W <- sqrt(2 * qf(level, 12, n - 12))
  X <- model.matrix(object, data = lobsters) # Design matrix
  ME <- rep(0,162)
  for (i in 1:n){
    dim(solve(t(X)%*%X))
    ME[i] <- W * sqrt(MSE * t(newdata[i])%*%(solve(t(X)%*%X))%*%newdata[i])
  }
  upr <- fit + ME
  return(upr)
}
whtrend <- function(mpa_type){
  trendpts<-pred_WH_upper(lob_dens_rootfit, newdata = cbind(MPA = rep(mpa_type,n),
                                                            new_data), level = 0.95)
  return(trendpts)
  }

plot2 +
  geom_smooth(aes(x=new_depthm, y = whtrend(mpa_type[1])),
                color = color[1]) +
  geom_smooth(aes(x=new_depthm, y = whtrend(mpa_type[2])),
                color = color[2]) +
  geom_smooth(aes(x=new_depthm, y = whtrend(mpa_type[3])),
                color = color[3]) +
```

```
  geom_smooth(aes(x=new_depthm, y = whtrend(mpa_type[4])),
                color = color[4]) +
  geom_smooth(aes(x=new_depthm, y = whtrend(mpa_type[5])),
                color = color[5])
```

(10) [3 pts] Fit either of the two standardized regression models discussed in class (either the one from the text, or the one that only standardizes the predictors) for the same cube-root response and set of predictors. Which variable is estimated to have the largest effect on lobster density for a change of 1 sample standard deviation in its respective predictor value

The largest effects by B_k value are Flat_Rock, Sand, and Boulder predictors.

```
X_alt <- scale(X[, -1])
fit_std <- lm(Lob_dens^(1/3) ~ X_alt,
              data = lobsters)
coef(fit_std)
```

```
##                                          (Intercept)
##                                          0.241771405
##            X_altMPALaguna Beach State Marine Reserve
##                                         -0.035096091
## X_altMPAPoint Vicente State Marine Conservation Area
##                                         -0.068406578
##          X_altMPASouth La Jolla State Marine Reserve
##                                          0.013770439
##       X_altMPASwami's State Marine Conservation Area
##                                         -0.010242728
##                            X_altInside_OutsideOutside
##                                          0.006718561
##                                          X_altDepth_m
##                                         -0.052900958
##                                        X_altRelief_cm
##                                         -0.003472298
##                                        X_altFlat_Rock
##                                          0.596039285
##                                          X_altCobble
##                                          0.210809881
##                                         X_altBoulder
##                                          0.451385296
##                                            X_altSand
##                                          0.479440502
```

(11) [2 pts] Compute $SSR(X_{\text{MPA}}|\dots)$, the extra regression sums of squares associated with including all levels of the categorical predictor MPA after including all other variables in the regression model for cube-root of lobster density.

The Extra Sum of Squares associated with the MPA predictor is 0.77419.

```
# Type III
car::Anova(lm(as.formula(paste("Lob_dens^(1/3)", pred_formula_str)), data = lobsters), type = 'III')
```

```
## Anova Table (Type III tests)
##
## Response: Lob_dens^(1/3)
##                 Sum Sq  Df F value    Pr(>F)
## (Intercept)    0.04060   1  1.9782  0.161653
## MPA            0.77419   4  9.4300 7.935e-07 ***
## Inside_Outside 0.00665   1  0.3241  0.570001
## Depth_m        0.29399   1 14.3238  0.000222 ***
## Relief_cm      0.00156   1  0.0758  0.783383
## Flat_Rock      0.06882   1  3.3529  0.069070 .
## Cobble         0.06386   1  3.1115  0.079777 .
## Boulder        0.08162   1  3.9768  0.047944 *
## Sand           0.05889   1  2.8692  0.092364 .
## Residuals      3.07869 150
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(12) [2 pts] Conduct an appropriate test to quantify the evidence against the null hypothesis that MPA is not linearly associated with the cube-root of lobster density after including all other predictors. Report the test statistic and p-value. Do you reject the null hypothesis at a level $\alpha = 0.01$?

The Type III anova test uses F* = (SSR(X_{MPA}/1) / (SSE(F)/(n-12)). The null hypothesis states that F* $\sim$ F (1,150) since we have n = 162, p = 12 The p-value is 7.935e-07 which is sufficient to reject the Null Hypothesis. We get this p-value from the previous type III Anova() code block.