HW5: Quantitative and Qualitative Predictors (ALRM Ch. 8)

MATH/STAT 571A

DUE: 11/3/2023 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 [20 pts]

(1) [6 pts] We have already seen that, for any predictor variable, X, the centered predictor $x = X - \bar{X}$ is orthogonal to the intercept, i.e., $x'1_n = 0$, where 1_n denotes the n-vector of 1s. We can use ideas from linear algebra to create a transformed quadratic predictor that is orthogonal to both the intercept and x. The idea is to begin with X^2, then "remove" the part that is not orthogonal to the other predictors. Recall from class: (i) how to construct a projection/hat matrix, $\mathbf{H}$, (ii) that pre-multiplying a vector by $\mathbf{H}$ projects the vector onto the linear space associated with $\mathbf{H}$, and (iii) that multiplying by $\mathbf{I} - \mathbf{H}$ projects a vector onto the orthogonal, residual space.

a. [3 pts] Let $\mathbf{X}_{n\times} = \begin{bmatrix} 1_n & x \end{bmatrix} = \begin{bmatrix} 1 & X_1 - \bar{X} \\ 1 & X_2 - \bar{X} \\ \vdots & \vdots \\ 1 & X_n - \bar{X} \end{bmatrix}$ be the matrix containing the intercept and centered first-order term, and let $X^2 = \begin{bmatrix} X_1^2 \\ \vdots \\ X_m^2 \end{bmatrix}$. Show that, for $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, the transformed vector $x_2 = (\mathbf{I} - \mathbf{H})\, X^2$ is orthogonal to both the intercept and x.

b. [3 pts] Use R to create a predictor vector of length $n = 10$ with any finite values you like, as long as $\sum (X_i - \bar{X})^2 > 0$. Create the model matrix $\begin{bmatrix} 1_n & x & x_2 \end{bmatrix}_{n\times 3}$ using the approach outlined above, and show, computationally, that x_2 is orthogonal to both 1_n and x.

```
x <- rnorm(10)
x_scaled <- scale(x)
X <- cbind(rep(1,10),x_scaled)
H <- X%*%solve(t(X)%*%X)%*%t(X)
I <- diag(10)
```

```
X_sq <- x^2
x_sq <- (I-H)%*%X_sq
paste("Results:", t(x_sq)%*%rep(1,10), t(x_sq)%*%x,sep=" ")
```

```
## [1] "Results: -2.22044604925031e-16 5.55111512312578e-17"
```

(2) [4 pts] (ALRM 8.18) Refer to regression model (8.49). Portray graphically the response curves for this model if $\beta_0 = 25$, $\beta_1 = 0.30$, $\beta_2 = -12.5$, and $\beta_3 = 0.05$. Describe the nature of the interaction effect.

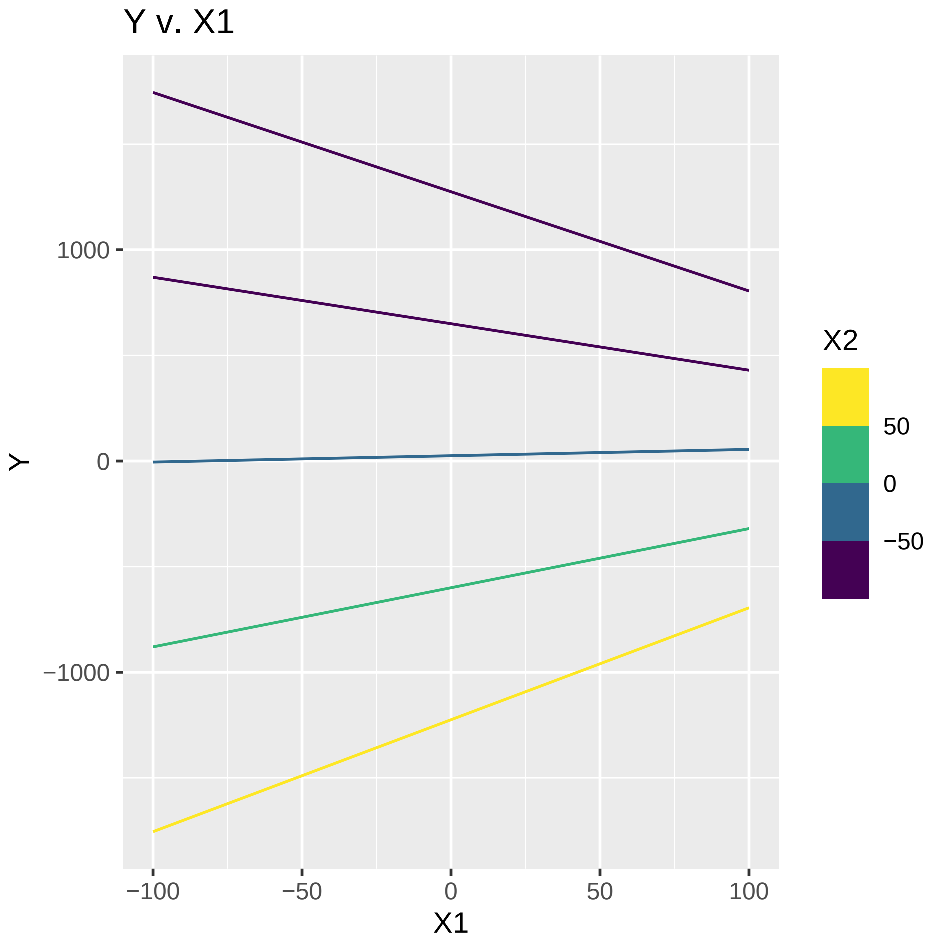
The interaction has a reinforcing effect on X1 and and interfering effect on X2. Y = B0 + X1 (B1 + B3 * X2) + X2 * B2 X1 always has a positive effect which increases as X2 increases and decreases as X2 decreases: (B1 + B3*X2) Y = B0 + X2 (B2 + B3* X2) + X1 * B1 X2 has a very negative effect but its effect becomes less negative as X1 increase and more negative as X1 decreases. (B2 + B3*X2)

In the graph, as X1's positive effect on Y decreases as X2 decreases.

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
B0 <- 25
B1 <- 0.3
B2 <- -12.5
B3 <- 0.05

X1_grid <- seq(-100,100,200/100)
X2_grid <- quantile(seq(-100,100), probs = seq(0, 1, 0.25))
newdata <- tibble(expand.grid(X1 = X1_grid, X2 = X2_grid))
newdata <- newdata %>% mutate(Y = (B0 + B1*X1 + B2*X2 + B3*X1*X2))
ggplot() +
  geom_line(aes(x = X1, y = Y, group = X2, col = X2), data = newdata) +
  scale_color_viridis_b() +
  ggtitle("Y v. X1")
```



(3) [4 pts] (ALRM 8.31)

a. [2 pts] Derive the expressions for b_0', b_1', and b_{11}' in (8.12a-c).

b. [2 pts] Using (5.46), obtain $\text{Var}(\mathbf{b}')$, the variance-covariance matrix for the regression coefficients, $\mathbf{b}'$, pertaining to the original X variable in terms of $\text{Var}(\mathbf{b})$, the variance-covariance matrix for the regression coefficients pertaining to $x = X - \bar{X}$ (*hint: start by writing* $\mathbf{b}' = \mathbf{Ab}$ *for a known matrix* $\mathbf{A}$).

(4) [6 pts] For two vectors X_1, X_2 of shared length n, let $X_1 \perp X_2$ denote orthogonality such that $X_1' X_2 = \sum_{i=1}^n X_{i1} X_{i2} = 0$, and let $X_1 X_2$ denote the vector resulting from pointwise multiplication as usual (i.e., $(X_1 X_2)_i = X_{i1} X_{i2}$). For each statement below, either prove it is true or demonstrate it is false with a counter-example:

a. [2 pts] $X_1 \perp X_2 \implies X_1 \perp X_1 X_2$.

b. [2 pts] $X_1 \perp X_2, X_1 \perp X_3,$ and X_2 not a scalar multiple of $X_3 \implies X_1 \perp X_2 X_3$.

c. [2 pts] $X_1 \perp X_2, X_1^2 \perp X_2 \implies X_1 \perp X_1 X_2$.

II. Lobsters in Southern California [20 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) [5 pts] Fit a linear regression model for the cube-root of lobster density using the predictors "MPA", "Depth_m", "Relief_cm", "Flat_Rock", "Cobble", "Boulder", "Sand" and the interaction between "Depth_m" and "Flat_Rock". Verify that the p-value associated with the main effect of "Depth_m" is 0.47138. Report the associated coefficient estimates and p-values for both main effects and the interaction. Given the large p-value associated with the effect of "Depth_m", should we consider removing it from the model? Why/why not?

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

fit_cube <- lm(Lob_dens^(1/3) ~ MPA + Relief_cm + Depth_m*Flat_Rock + Cobble + Boulder +
    Sand, data = lobsters_subset)
summary(fit_cube)
```

```
##
## Call:
## lm(formula = Lob_dens^(1/3) ~ MPA + Relief_cm + Depth_m * Flat_Rock +
##     Cobble + Boulder + Sand, data = lobsters_subset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32397 -0.08230  0.00859  0.09415  0.43445
##
## Coefficients:
##                                                   Estimate Std. Error t value
## (Intercept)                                     -1.454e+00  9.725e-01  -1.495
## MPALaguna Beach State Marine Reserve            -1.030e-01  3.761e-02  -2.738
## MPAPoint Vicente State Marine Conservation Area -2.644e-01  4.898e-02  -5.397
## MPASouth La Jolla State Marine Reserve           1.361e-02  3.619e-02   0.376
## MPASwami’s State Marine Conservation Area       -5.029e-02  3.485e-02  -1.443
## Relief_cm                                       -6.416e-05  2.839e-04  -0.226
## Depth_m                                         -4.293e-03  5.946e-03  -0.722
## Flat_Rock                                        2.051e-02  9.802e-03   2.093
## Cobble                                           1.703e-02  9.780e-03   1.741
## Boulder                                          1.988e-02  9.861e-03   2.016
## Sand                                             1.683e-02  9.743e-03   1.727
## Depth_m:Flat_Rock                               -2.716e-04  1.079e-04  -2.517
##                                                 Pr(>|t|)
## (Intercept)                                      0.13709
## MPALaguna Beach State Marine Reserve             0.00693 **
## MPAPoint Vicente State Marine Conservation Area 2.59e-07 ***
## MPASouth La Jolla State Marine Reserve           0.70734
## MPASwami’s State Marine Conservation Area        0.15107
## Relief_cm                                          0.82151
## Depth_m                                            0.47138
## Flat_Rock                                          0.03806 *
## Cobble                                             0.08375 .
## Boulder                                            0.04555 *
## Sand                                               0.08621 .
## Depth_m:Flat_Rock                                  0.01290 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1405 on 150 degrees of freedom
## Multiple R-squared:  0.3478, Adjusted R-squared:    0.3
## F-statistic: 7.272 on 11 and 150 DF,  p-value: 7.139e-10
```

(6) [5 pts] Give an interpretation for the estimated interaction effect in the context of this data set. For what combinations of depth and flat rock cover would we expect to see the most lobsters? Flat_rock has a positive effect. Since depth_m has a negative effect and decreases the positive effect of flat_rock, depth_m always has a negative effect. Then, less depth_m more flat_rock will lead to the highest lobster density.

(7) [5 pts] Fit a new version of the model with all the same predictors as in (6), but use centered and scaled versions of the continuous predictors. Which continuous predictor has the largest estimated effect? (*Hint: first create a new dataframe called* `lobsters_scaled` *equal to* `lobsters` *and then replace the appropriate columns in* `lobsters_scaled` *with modified vectors using the* `scale()` *function.*)

Flat Rock has the largest effect after scaling.

```
names(lobsters_subset)
```

```
## [1] "Lob_dens"   "MPA"       "Depth_m"   "Relief_cm" "Flat_Rock" "Cobble"
## [7] "Boulder"   "Sand"
```

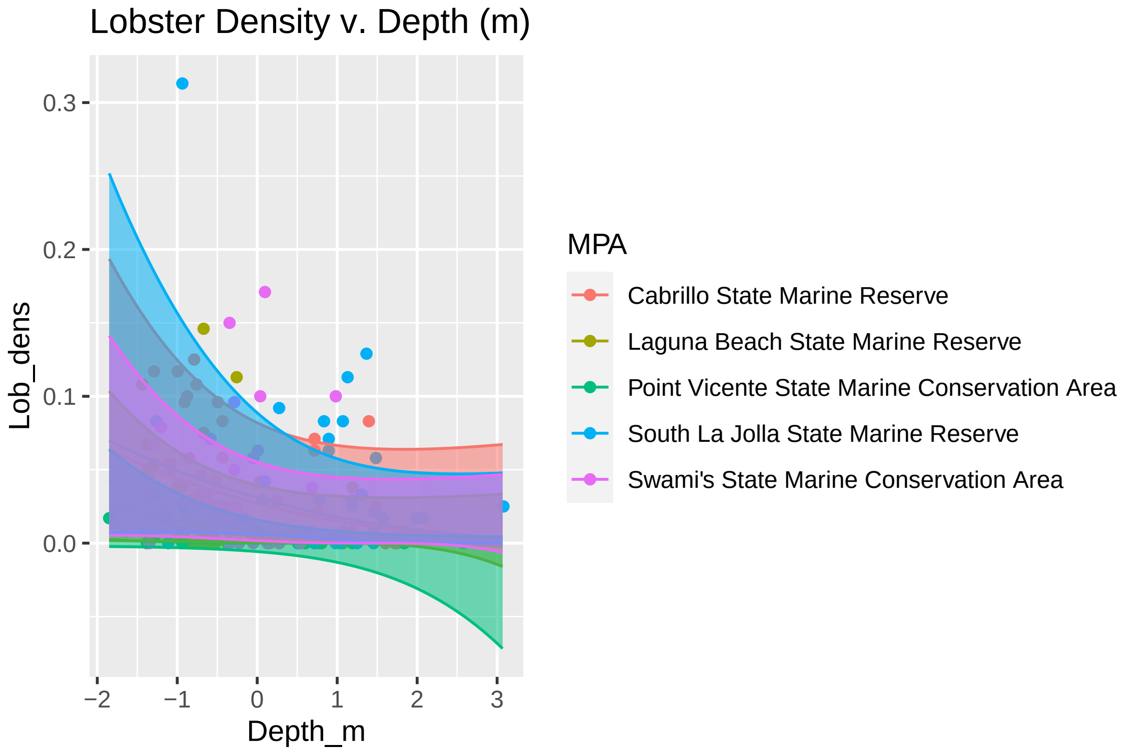
```
lobsters_scaled = data.frame(lobsters_subset[1:2], scale(lobsters_subset[,c(-1,-2)])) # scale cts predi
fit_std_cube <- lm(Lob_dens^(1/3) ~ MPA + Relief_cm + Depth_m*Flat_Rock + Cobble + Boulder +
    Sand, data = lobsters_scaled)
summary(fit_std_cube)
```

```
##
## Call:
## lm(formula = Lob_dens^(1/3) ~ MPA + Relief_cm + Depth_m * Flat_Rock +
##     Cobble + Boulder + Sand, data = lobsters_scaled)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32397 -0.08230  0.00859  0.09415  0.43445
##
## Coefficients:
##                                                 Estimate Std. Error t value
## (Intercept)                                     0.300345   0.025333  11.856
## MPALaguna Beach State Marine Reserve           -0.102969   0.037610  -2.738
## MPAPoint Vicente State Marine Conservation Area -0.264371   0.048981  -5.397
## MPASouth La Jolla State Marine Reserve          0.013614   0.036193   0.376
## MPASwami's State Marine Conservation Area      -0.050294   0.034851  -1.443
## Relief_cm                                        -0.002779   0.012296  -0.226
## Depth_m                                          -0.052890   0.013607  -3.887
## Flat_Rock                                         0.590916   0.319176   1.851
## Cobble                                            0.204060   0.117216   1.741
## Boulder                                           0.447483   0.221927   2.016
## Sand                                              0.479193   0.277461   1.727
## Depth_m:Flat_Rock                                -0.030077   0.011951  -2.517
##                                                  Pr(>|t|)
## (Intercept)                                       < 2e-16 ***
## MPALaguna Beach State Marine Reserve             0.006933 **
## MPAPoint Vicente State Marine Conservation Area 2.59e-07 ***
## MPASouth La Jolla State Marine Reserve           0.707339
## MPASwami's State Marine Conservation Area        0.151073
## Relief_cm                                        0.821505
## Depth_m                                          0.000152 ***
## Flat_Rock                                        0.066081 .
## Cobble                                           0.083754 .
## Boulder                                          0.045548 *
## Sand                                             0.086214 .
## Depth_m:Flat_Rock                                0.012899 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1405 on 150 degrees of freedom
## Multiple R-squared:  0.3478, Adjusted R-squared:    0.3
## F-statistic: 7.272 on 11 and 150 DF,  p-value: 7.139e-10
```

(8) [5 pts] Use either the model from (5) or the scaled version in (7) to produce the following figure:

a. [1 pt] Plot the observed density of lobsters as a function of observed depth (or centered and scaled depth) and color each point based on its associated MPA. Include a legend.
b. [2 pts] For each MPA, add a curve that shows the predicted lobster density (***not*** *the cube-root of lobster density*) over a grid of depths, with all other variables set to their respective sample means. Match the colors of the curves to the colors used for the points.
c. [2 pts] Add a polygon/ribbon of the appropriate color around each predicted curve that shows the associated 95% Working-Hotelling confidence band.

```r
# using version (7) centered/scaled
library(ggplot2)
pred_WH_MLR <- function(object, newdata, level = 0.95){
  n <- nrow(object$model) ## extract n from model object
  p <- object$rank

  # PW ci
  fit <- predict(object, newdata, interval = 'confidence')
  ME_confidence <- fit[,'fit'] - fit[,'lwr']

  # rescale to WH ci
  alpha <- 1-level
  t_alpha <- qt(1-alpha/2, df = n-p)
  W_alpha <- sqrt(p * qf(level, n-p, p))

  ME <- ME_confidence*W_alpha/t_alpha
  upr <- fit[,'fit'] + ME
  lwr <- fit[,'fit'] - ME
  return(cbind(fit[,'fit'], lwr, upr))
}

depthMgrid <- seq(min(lobsters_scaled$Depth_m),max(lobsters_scaled$Depth_m),1/25)
newdata <- expand.grid(Depth_m = depthMgrid,
            Relief_cm = mean(lobsters_scaled$Relief_cm),
            Flat_Rock = mean(lobsters_scaled$Flat_Rock),
            Cobble = mean(lobsters_scaled$Cobble),
            Boulder = mean(lobsters_scaled$Boulder),
            Sand = mean(lobsters_scaled$Sand),
            MPA = unique(lobsters_scaled$MPA))
predictions <- as.data.frame(cbind(newdata,pred_WH_MLR(fit_std_cube,newdata=newdata)^3))
colnames(predictions)[8] <- "fit"


ggplot(data = lobsters_scaled) +
  geom_point(aes(x=Depth_m,y=Lob_dens,color=MPA)) +
  geom_line(aes(x=Depth_m,y=fit,color=MPA), data = predictions) +
  geom_ribbon(aes(x = Depth_m, ymin = lwr, ymax = upr, alpha = .1, color = MPA, fill = MPA),
              data = predictions, show.legend=F) +
  ggtitle("Lobster Density v. Depth (m)")
```


Lobster Density v. Depth (m)
Lob_dens
Depth_m
MPA
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