Lab Report3 Appendix

11/01/2023

- Import Library: alr3, reshape2, GGally, tidyverse, car, RNOmni
- loaded dataset: dataset_Facebook.csv

Exploratory Data Analysis

```
missing_values <- sapply(df, function(x) sum(is.na(x)))
print("Summary of Missing Values:")
print(missing_values[missing_values > 0])
data.frame(do.call(cbind, lapply(na.omit(df), summary)))
```

```
colmapping <- letters[0:ncol(df)]
head(tibble(colnames(df), colmapping))
mapped_df <- df %>% rename(!!!setNames(names(.), colmapping))
head(mapped_df)
```

```
numerical_df <- mapped_df[sapply(mapped_df, is.numeric)]
corr_mat <- cor(numerical_df, use = "complete.obs")

melted_corr <- melt(corr_mat)
ggplot(data = melted_corr, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(
    low = "blue",
    high = "red",
    mid = "white",
    midpoint = 0,
    limit = c(-1, 1),
    space = "Lab",
    name = "Pearson\nCorrelation"
  ) +
  scale_x_discrete(expand = c(0, 0)) +
  scale_y_discrete(expand = c(0, 0)) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 0)) +
  ggtitle("Correlation Matrix of Continuous Variables")
```

```
num_df <- na.omit(numerical_df)
ggpairs(num_df, progress = FALSE,
        upper = list(continuous = wrap("cor", size = 1)),
        lower = list(continuous = wrap("points", alpha = 0.3,    size=0.05)),
        title = "Scatterplot matrix of Facebook Metrics") +
```

```
  theme(axis.line=element_blank(),
        axis.text=element_blank(),
        axis.ticks=element_blank())
```

Statistical Modeling

1st Question of Research Interest: The effect of paid promotions on visibility of Facebook posts

```
mode_paid <- as.numeric(names(sort(table(df$Paid), decreasing=TRUE)[1]))
df$Paid[is.na(df$Paid)] <- mode_paid

df$Lifetime.Post.Total.Reach[is.na(df$Lifetime.Post.Total.Reach)] <- median(df$Lifetime.Post.Total.Reach

df$Lifetime.Post.reach.by.people.who.like.your.Page[is.na(df$Lifetime.Post.reach.by.people.who.like.you

df$Total.Interactions[is.na(df$Total.Interactions)] <- median(df$Total.Interactions, na.rm=TRUE)

df$Page.total.likes[is.na(df$Page.total.likes)] <- median(df$Page.total.likes, na.rm=TRUE)

with(df, plot(Lifetime.Post.reach.by.people.who.like.your.Page, Lifetime.Post.Total.Reach,
              main="Scatter Plot",
              xlab="Lifetime.Post.reach.by.people.who.like.your.Page",
              ylab="Lifetime.Post.Total.Reach"))
```

```
with(df, plot(Total.Interactions, Lifetime.Post.Total.Reach,
              main="Scatter Plot",
              xlab="Total.Interactions",
              ylab="Lifetime.Post.Total.Reach"))
```

```
with(df, plot(Page.total.likes, Lifetime.Post.Total.Reach,
              main="Scatter Plot",
              xlab="Page.total.likes",
              ylab="Lifetime.Post.Total.Reach"))
```

Initital Multiple Linear Regression

```
initial_mlr_model <- lm(Lifetime.Post.Total.Reach ~ Paid + Lifetime.Post.reach.by.people.who.like.your.

summary(initial_mlr_model)
```

Initial MLR disgnostics

```
par(mfrow=c(2,2))
plot(initial_mlr_model, which=1:3)

avPlots(initial_mlr_model)
```

Log-log transformed Model (fm)

```
df$Total.Interactions <- df$Total.Interactions + 1e-10 # To avoid log(0)=-inf
fm <- lm(log(Lifetime.Post.Total.Reach) ~ Paid + log(Lifetime.Post.reach.by.people.who.like.your.Page) +
            data=df)
summary(fm)
```

Handle Bad High Leverage Point

There is no bad leverage point in the model

```
hat_values <- hatvalues(fm)
std_residuals <- rstandard(fm)
p = 3
n = 500
high_leverage <- which(hat_values > (2*(p+1))/n)
outliers <- which(abs(std_residuals) > 2)

to_remove <- intersect(high_leverage, outliers)
length(to_remove)
```

log-log transformed Model Diagnostics

```
par(mfrow=c(2,2))
plot(fm, which=1:3)

## Log-log transformed Model AV Plots
avPlots(fm)
```

2ed Question of Research Interest: The association between page total likes and types of post

```
attach(df)
table(df$Type, useNA = 'always')
boxplot(Page.total.likes ~ Type)
```

```
boxplot(Page.total.likes ~ Category)
```

```
boxplot(Page.total.likes ~ Paid)
```

```
ln_like <- log(Page.total.likes)
hist(ln_like)
```

```
hist(Page.total.likes)
```

```r
tapply(Page.total.likes,Type, summary)
df$rn_page.total.likes <- RankNorm(Page.total.likes,ties.method = "average")
summary(df$rn_page.total.likes)
sd(df$rn_page.total.likes)
hist(df$rn_page.total.likes)
```

```r
inv <- 1/Page.total.likes
hist(inv)
```

```r
minmax <- (Page.total.likes-min(Page.total.likes))/(max(Page.total.likes)-min(Page.total.likes))
hist(minmax)
```

```r
attach(df)
```

```
## The following objects are masked from df (pos = 3):
##
##     Category, comment, Lifetime.Engaged.Users,
##     Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post,
##     Lifetime.Post.Consumers, Lifetime.Post.Consumptions,
##     Lifetime.Post.Impressions.by.people.who.have.liked.your.Page,
##     Lifetime.Post.reach.by.people.who.like.your.Page,
##     Lifetime.Post.Total.Impressions, Lifetime.Post.Total.Reach, like,
##     Page.total.likes, Paid, Post.Hour, Post.Month, Post.Weekday, share,
##     Total.Interactions, Type
```

```r
## Page.total.likes highly skewed toward the left - need to transform but first proceed with the MV mod
## Inital variable selection from scatter plot + heatmap
## Type Post.Month Post.Weekday Post.Hour Total.Interactions Lifetime.Post.Consumers

plot(Page.total.likes ~ Post.Month)
```

```r
plot(Page.total.likes ~ Post.Weekday)
```

```r
plot(Page.total.likes ~ Post.Hour)
```

```r
plot(Page.total.likes ~ like)
```

```r
plot(Page.total.likes ~ share)
```

```r
plot(Page.total.likes ~ comment)
```

```r
plot(Page.total.likes ~ Total.Interactions,subset = Total.Interactions < 6000)
```

```r
plot(Page.total.likes ~ Lifetime.Post.Consumers)
```

Fit the MV model

```r
## dummy variable for type
df$type.photo <- ifelse(Type == c("Photo"),1,0)
df$type.status <- ifelse(Type == c("Status"),1,0)
df$type.video <- ifelse(Type == c("Video"),1,0)

## simple linear model

m00 <- lm(Page.total.likes~type.photo+type.status+type.video, data=df)
summary(m00)
plot(m00)
```

```r
## simple linear model - rank normalized
m0 <- lm(rn_page.total.likes~type.photo+type.status+type.video, data=df)
summary(m0)
plot(m0)
```

```r
## first model - initial try
m1 <- lm(rn_page.total.likes~type.photo+type.status+type.video+Total.Interactions+Lifetime.Post.Consume
summary(m1)
plot(m1)
```

```r
AIC(m1)

## outlines diagnostics are way off... will need to: 1) remove bad outliers
##                                                   2) add in a squared term
##                                                   3) check collinearity


## better but still not good - remove bad outliers 447 245
fb2 <- df[-c(447,245),]
m2 <- lm(rn_page.total.likes~type.photo+type.status+type.video+Total.Interactions+Lifetime.Post.Consume
summary(m2)
plot(m2)
```

```r
AIC(m2)

##remove consumption
m3 <- lm(rn_page.total.likes~type.photo+type.status+type.video+Total.Interactions+Lifetime.Post.Consume
summary(m3)
plot(m3)
```

```r
##partial F test
anova(m2,m3)

## check collinearity

vif(m3)

## final model

summary(m3)
```

Model Diagnostic

```
## m3
par(mfrow=c(2,2))
plot(m3)
```

```
avPlots(m3)
```

3rd Question of Research Interest: The association between Post Consumption and unique users engagement

```
q1<-df %>% select(Lifetime.Post.Consumptions,Lifetime.Post.Consumers,Lifetime.Engaged.Users,Lifetime.Pe
newname<-c("consumptions", "consumers", "engagedu", "likeandengaged")
tibble(colnames(q1),newname)
mq1<-q1 %>% rename(!!!set_names(names(.), newname))

q2<- df %>% select(Lifetime.Post.Consumptions,Lifetime.Engaged.Users)
mq2<-q2 %>% rename(!!!set_names(names(.), newname[c(1,3)]))
```

Initial Model

Despite the F-value shows that the model is significant, only one predictor is significant. There is a clear pattern in the standardized residual plot despite the high p-value indicating the significance of predictors. Therefore, the model is invalid.

```
model1<-lm( consumptions ~ consumers + engagedu + likeandengaged , data = mq1)
summary(model1)
pairs(consumptions~consumers + engagedu + likeandengaged, data = mq1)
```

```
summary(bctrans(model1))
```

```
## Warning: 'bctrans' is deprecated.
## Use 'powerTransform' instead.
## See help("Deprecated") and help("alr3-deprecated").
```

log-log transform and Diagnostic Plot

```
pairs(log(consumptions)~log(consumers) + log(engagedu) + log(likeandengaged) , data = mq1)
```

```
model11<-lm(log(consumptions)~log(consumers) + log(engagedu) + log(likeandengaged) , data = mq1)
summary(model11)
par(mfrow=c(2,2))
plot(model11)
```

```r
par(mfrow=c(2,2))
avPlots(model11)
```