# Step 1: Make sure data are clean

summary(retailer)

# remove columns/rows, address NAs, convert types, if needed #catalog has only 4 so it shud be a factor #suspect in categorical values is few levels having too few observation or no values # shud remove

# Step 2: Examine at a univariate level (transform if needed)

# factors have reasonable levels and distribution #all other factors have only 3 levels #all levels have even distribution of values

* Remove outliers, powertransform , 4 plots pick the best
* Try the simplest forms, i.e., log, sqrt
* # Try creating an index that consists of multiple highly related variables
* # Try creating a factor out of "bad" variables

# Step 3: Examine bivariate relations (transform if needed)

# Which IVs do we need to included # Explore correlation between DV and IVs, and among IVs

# Now DV and IVs, also consider relationships among IVs

# A correlation plot would be useful, but we only have two numeric variables in the data

# Let's explore other relationships in the data

cor.test(retailer$expense,retailer$income)

summary(aov(expense ~ agegroup, data=retailer))

chisq.test(retailer$catalog, retailer$agegroup)

# Step 4: modeling

mod1 <- lm(sqrt(expense)~income+agegroup+gender+homeownership+maritalstatus+location, data=retailer)

include ones which were statistically significant

interpret results

#Interpretation:

*#Impressions are statistically highly significant(positive relation) and gender male, gender unknown compared to female are statistically highly significant(negative realtion)*

*#One unit increase of Impression effects in .00014 units increase of sqrt of purchases*

*#One unit increase of gender-male effects in .9495 units decrease in sqrt of purchases compared to gender-female*

*#One unit increase of gender-unknown effects in 2.04 units decrease in sqrt of purchases compared to gender-female*

*summary(mod1)$r.squared*

*summary(mod1)$adj.r.squared #.829 #The IV variables are able to explain 82.9% of variance of sqrt of purchases*

summary(mod1)$adj.r.squared

#table (as.factor(real$bath))

# Interactions #same slope parallel on top of each other then no interaction #if diff slope then there is interaction is there a diff slope? if so yes. Here bigger lotisize houses is pref area increases their price soon. intercation effect. so wen both lotsize n pref area meets has highr slope

#prefarea n recroom has a slight sign interaction #when a house is in prefarea and has a recroom, the log price decreases by 15% as compared to not having both recroom and not being in prefarea

# Diagnostics

#=

plot(mod2) # This is a great way to identify multiple possible concerns with the model

# For example, outliers. #this gives outlier in bivariate #gotta remove #in all 4 plots the outliers reapats we gotta remove them

# Better yet:

par(mfrow=c(2,2)) # This function sets up the plotting canvas to 2x2, so that we don't need to go through them one by one. Now again:

plot(mod2)

Step1;

summary(retailer)

step2:

View(describe(retailer[,c("income", "expense")])) #talk abt skew n spread #expense median 944 sd also similar so this has wider spread than first

plot(density(retailer$income)) # not perfect

plot(density(retailer$expense))

boxplot(retailer$expense) # quite a few outliers

outliers(retailer$expense) # only a few are extreme outliers #no extreme outlier now

retailer <- retailer[-c(48,193,448,497,603,675),]

expense – .25 income – sqrt

# Build a model to predict house prices

Plot price n see if skwed or not – take plot log or so

#fr bathrms he wants to c if it is factor#convert #mod log cuz log helped #

#table (as.factor(real$bath)) #u see 4 bthm - 1 3-10

# Find a model that has the best adj-r-squared – model first add few, then change n add few calc r sq at each leavel

# Know how to interpret the results!

box1 <- boxoffice\_cln[boxoffice\_cln$rating == "PG" |boxoffice\_cln$rating == "R" | boxoffice\_cln$rating == "PG-13", ]

View(box1)

#dropping extra column created by me

boxoffice\_cln <- boxoffice\_cln[,c(-7)]

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 11.904414455 0.032685618 364.210 < 0.0000000000000002 \*\*\*

lotsize 0.000076522 0.000006188 12.366 < 0.0000000000000002 \*\*\*

prefareayes 0.245654365 0.035318880 6.955 0.0000000000112 \*\*\*

recroomyes 0.230165495 0.045268296 5.084 0.0000005238102 \*\*\*

prefareayes:recroomyes -0.158853306 0.072182631 -2.201 0.0282 \*