**Interpretation:**

For every dollar of income the odds of default go down by 0.007(.993). not percentage

For those who went to college the odds of default go up by 87%(1.875) Number multiplied by 1.875

taking a loan multiplies the odds of y by about .64,compared to those who did not take a loan, holding the other IVs at their constant

**factor analysis:**

Hip factor, value proposition, quality proposition

Aesthetics, functional benefits, economic value, credibility

<https://www.promptcloud.com/blog/exploratory-factor-analysis-in-r/>

Looking at this plot and parallel analysis, anywhere between 2 to 5 factors factors would be good choice.

Results suggest the optimal number of factors based on the data. Look for an "elbow", especially around eigenvalue = 1.

**AIC:** Index of fit that penalizes the number of parameters. So, the smaller the AIC, the better. We can compare models using AIC but we don't get a good sense of our model's performance . Instead, we use pseudo-R-Squared measures. Pseudo r sq bigger better. A pseudo R-squared only has meaning when compared to another pseudo R-squared of the same type, on the same data, predicting the same outcome. In this situation, **the higher pseudo R-squared indicates which model better predicts the outcome. Better adj r sq better. If adding variable adj r sq goes down means collinearity smthn**

deviance /residul the value is less than 1 so no **overdispersion**

But how do we know for sure? We compare the model to a model that assumes overdispersion, and see if there is a difference

If I do indeed have overdispersion i need to then run a quasi-binomial model, if chi-sq value is above p –value then the test failed to find any problems.

p-value is much higher than 0.05 # There is no overdispersion.#if value is less than .05 then there is overdispersion #anything above std p-value no overdispersion

12th session:

The problem: identify and remove outliers. # How many outliers are in the data?

# The problem: can you start with a regression model then perform EDA? Not recommnded. – do uni bi n then model - Look at the z-value for poutcome success. Useful fr prediction ?too few levels? Recode

Remember that you can recode a variable with an ifelse function

Can't transform a variable. Getting an error message

bank$balance[bank$balance<=0] %>% length/nrow(bank)

bank$balance\_std <- bank$balance+abs(min(bank$balance))+1

# Now try the transformation

log(bank$balance\_std)

Logistic diagnostics:

Examine for possible multicollinearity

Check fr overdispersion,

Model fit:

AIC, Pseudo r2

Remove one IV n keep checking

Factor analysis:

#1. Examine the data

#2. Scale the data (if needed)

d1 <- data.frame(scale(decathlon[,1:10]))

#3. Consider number of factors

#4. Extract factors

#5. (optional) use the factors discovered in analysis and modeling

SS loading higher than .7

Loading above .5 take.. take only + or only –

prop var --.206 factor 1 explains 20% of variance

keep all IV which are factors together and other var separately and do analysis

#exlore interactions it may matter. if tme was der will explore interactions

**Main Logistic steps**

## Step 1. Make sure the data are clean #is.na

# inspect the data

head(bank)

tail(bank)

bank %>% sample\_n(20) %>% View()

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

#=

# Outcome variable

tab <- table(bank$y)

#table(bank$y) %>%prop.table %>% round(2)

barplot(tab, col=c("orange", "steelblue"), main="Accepted Offer")

# Predictor variables

# this step is to select relevant IVs and prepare them as needed

str(bank)

## Step 3: Examine bivariate relations with relevant IVs and DV

# Default might be relevant

table(bank$default) # very few default cases

# still, is there a relationship?

xt\_def\_y <- xtabs(~y+default, data=bank) #xtabs fr header

summary(xt\_def\_y) # chisq.test; nope

# Education?

xt\_educ\_y <- xtabs(~y+education, data=bank)

summary(xt\_educ\_y) # chisq.test; better!

# We also want to explore a balance (numeric)

plot(density(bank$balance))

boxplot(bank$balance~bank$y,

main="Relationship between balance and offer acceptance",

col=rainbow(2))

# better #wo outliers #from the boxplot we see tat ppl who accept offer have higher balance

## Step 4. Modeling

bank\_mod1 <- glm(y~education+balance, data=bank1, family=binomial) #binomial approximates logit

#glm(y~education+balance, data=bank1, family=binomial) %>% summary #if u dont want to save model n just view

summary(bank\_mod1)

exp(coef(bank\_mod1))

confidence intervals for predictions

exp(confint(bank\_mod1, level=.99)) # default is .95 for those who took a loan, the odds multiply between .3 and .7 compare to those who did not take a loan.

# Add more variables

bank\_mod2 <- glm(y~balance+education+age+marital+loan+poutcome,

data=bank1, family=binomial)

summary(bank\_mod2)

# Second, test the difference

anova(bank\_mod1, bank\_mod2, test="Chisq")

# Now re-run the previous procedure

exp(coef(bank\_mod2))

**# Predict outcome #we r going to get odds so ll be betn 0 n 1**

set.seed(123)

samp <- sample(1:nrow(bank), 100)

pred <- predict(bank\_mod3,

bank[samp, c("balance", "marital", "loan","poutcome")],

type="response") %>% round(1)

ylikely <- which(pred>0.5) %>% names %>% as.numeric

bank[ylikely,c("balance", "marital", "loan", "poutcome", "y")] %>% View

# Interactions? # very much like OLS

bank\_mod4 <- glm(y~marital+loan+balance\*poutcome, data=bank, family=binomial)

summary(bank\_mod4)

plot(effect(term="jobsatisfaction:age", mod=quitmod, default.levels=20), multiline=T)

5. Diagnostics

Examine for possible multicollinearity

vif(quitmod)

sqrt(vif(quitmod))>2

Check fr overdispersion,

Model fit:

AIC, Pseudo r2

Remove one IV n keep checking