**[Question no. 1]**

Variables chosen:

* 1. Respondent’s Income (RINCOME),
  2. How many in family earned money (EARNRS),
  3. Whether respondent is self-employed or works for somebody (WRKSLF),
  4. Satisfied with financial situation (SATFIN).

**Code:**

source('GSS-CleanData.R')

########## Clean Data #################

# Respondent's Income (RINCOME)

GSS.data$RINCOME[GSS.data$RINCOME==0 | GSS.data$RINCOME > 12] <- NA

# How many in family earned money (EARNRS)

GSS.data$EARNRS[GSS.data$EARNRS==9] <- NA

# Self-employed (WRKSLF)

GSS.data$WRKSLF[GSS.data$WRKSLF==0 | GSS.data$WRKSLF>7] <- NA

# Satisfied with financial situation (SATFIN)

GSS.data$SATFIN[GSS.data$SATFIN==0 | GSS.data$SATFIN>3] <- NA

GSS.data$SATFIN <- 4-GSS.data$SATFIN

# reduce GSS data to cases with responses in all relevant variables.

GSS.data.2 <- GSS.data[complete.cases(GSS.data[,c("JobSat", "RINCOME", "EARNRS","SATFIN", "WRKSLF")]),]

**[Question no. 2]**

Respondent’s Income: The relationship between income and job satisfaction is the subject of a vast body of research. While there is some disagreement on the strength of this relationship, there is at least a marginal relationship between the two (Judge, 2010). Therefore, I expect to find a positive relationship between income and job satisfaction because higher incomes may lead to a better quality of life style which might positively impact the job satisfaction level. I expect the effect size to be moderate, around 0.4, since even though income is an important variable, it is present in the data on a 10-point scale, reducing the effect of a single unit change in income.

How many people in the family earned money: In their study of job satisfaction in the Chinese workforce, Nielsen and Smyth found that married employees are more likely to have higher job satisfaction (Nielsen & Smyth, 2008). One of the possible reasons they proposed was that partners can rely on each other’s income and so can take jobs that are more satisfying even though they pay less. We build upon this line of reasoning by studying the relationship between the number of members in a family who are earning and job satisfaction. I expect to find an increase in job satisfaction as the number of members in families who are earning increases. I expect the effect size to be small, about 0.2, since there is no strong literature to support the existence of a relationship.

Whether the respondent is self-employed or works for somebody: Popular culture in recent times tends to glorify entrepreneurs, portraying them as being more ambitious and driven than the ordinary employee. Entrepreneurs also operate with a large degree of autonomy and stand to profit directly from their work. Not only does it seem natural for this to result in higher job satisfaction, previous works such as Tchinda and Bing’s 2017 study of data of Cameroon data also supports the existence of a relationship between the two (Tchinda & Bing, 2017). Therefore, I expect to find that self-employed respondents have higher job satisfaction than those who work for others. I expect the effect size to be relatively high, around 0.4, since although this is likely not the most important variable, its binary nature means that the change of a single unit in this variable will have a notable effect on job satisfaction.

Satisfied with the financial situation: Finances play an undeniably important role in our lives. I hypothesize that employees are unlikely to be satisfied with their jobs if their financial situation is unsatisfactory. I expect to find a strong positive relationship between the two. Given the qualitative nature of this variable as opposed to the quantitative nature of the SATFIN variable (one person may be satisfied by an income that is unsatisfactory for someone else), I expect this to be the most important variable and have a large effect size (0.5+).

**[Question no. 3]**

Null and alternative hypothesis in NHST format for SATFIN and JOBSAT:

The null hypothesis is that there is no significant relation between satisfaction in financial situation (SATFIN) and job satisfaction (JOBSAT) whereas the alternative hypothesis is that there is a significant relation between satisfaction in financial situation (SATFIN) and job satisfaction (JOBSAT).

Null and alternative hypothesis in conventional format for the SATFIN and JOBSAT:

Financial situation mediates the job satisfaction level such that employees are unlikely to be satisfied with their jobs if their financial situation is unsatisfactory whereas they are more likely to be satisfied with their jobs if their financial situation is satisfactory.

Hypothesis in Cumming’s format for the SATFIN and JOBSAT:

We aim to find, to what extent does satisfaction in financial situations impact the level of job satisfaction by reporting the regression coefficients along with their 95% confidence intervals.

**[Question no. 4]**

Hypothesis in Cumming’s format for the RINCOME and JOBSAT

We want to find to what degree does the respondent’s income affect the level of satisfaction in the job and report the regression coefficients along with their 95% confidence intervals.

Hypothesis in Cumming’s format for the EARNRS and JOBSAT

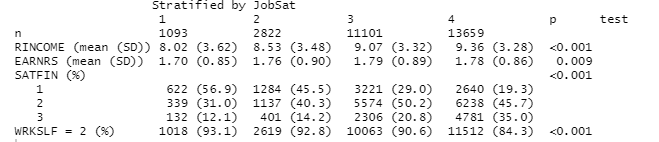
Our goal is to find out what is the effect of the number of earners in the family on the level of satisfaction in the job and report the coefficients of regression with the 95% confidence intervals.

Hypothesis in Cumming’s format for the WRKSLF and JOBSAT

The objective is to see to what is the effect of self employment on job satisfaction as opposed to being employed by someone else and also report the coefficients of regression with their 95% confidence intervals.

**[Question no. 5]**

Table 1



**Code: (Nishida, n.d.)**

#Table 1

library(tableone)

library(dplyr)

run\_tableone <- function(df, listVars, catVars, strata) {

x <- CreateTableOne(vars = listVars, data = df, factorVars = catVars, strata=strata)

as.data.frame(print(x)) %>%

tibble::rownames\_to\_column("Name")

}

View(run\_tableone(GSS.data.2,c("RINCOME", "EARNRS",

"SATFIN", "WRKSLF"),

c("SATFIN","WRKSLF"),c("JobSat")))

**[Question no. 6]**

**Code:**

######### Regression ##############

Model <- lm(JobSat ~ RINCOME + EARNRS + SATFIN + WRKSLF, data = GSS.data.2)

summary(Model)

confint(Model)

**R Output:**

Call:

lm(formula = JobSat ~ RINCOME + EARNRS + SATFIN + WRKSLF, data = GSS.data.2)

Residuals:

Min 1Q Median 3Q Max

-2.78587 -0.32138 0.01847 0.67862 1.09773

Coefficients:

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

(Intercept) 3.068116 0.033245 92.288 <2e-16 \*\*\*

RINCOME 0.015258 0.001378 11.069 <2e-16 \*\*\*

EARNRS 0.001482 0.005215 0.284 0.776

SATFIN 0.248307 0.006289 39.483 <2e-16 \*\*\*

WRKSLF -0.214704 0.013960 -15.380 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7692 on 28670 degrees of freedom

Multiple R-squared: 0.07012, Adjusted R-squared: 0.07

F-statistic: 540.5 on 4 and 28670 DF, p-value: < 2.2e-16

2.5 % 97.5 %

(Intercept) 3.002954623 3.13327807

RINCOME 0.012556211 0.01795971

EARNRS -0.008740111 0.01170340

SATFIN 0.235980334 0.26063335

WRKSLF -0.242066074 -0.18734245

**[Question no. 7]**

RINCOME, SATFIN and WRKSLF are significant predictors of job satisfaction since their p-value is extremely small (2.2e-16). EARNRS is not a significant predictor since the p-value is quite large (0.776). Furthermore, the 95% confidence intervals of RINCOME are from 0.012556211 to 0.01795971, for SATFIN the intervals are from 0.235980334 to 0.26063335 and for WRKSLF are from -0.242066074 to -0.18734245. All of these intervals do not contain the value 0 and hence the predictors corresponding to them are significant. The 95% confidence intervals of EARNRS are from -0.008740111 to 0.01170340 and contain the value 0 hence this variable is not significant.

The effect size or regression coefficient of RINCOME shows that upon increasing the respondent’s income by one unit an increase of 0.015 could be seen in the job satisfaction. The effect size of EARNRS shows that upon increasing the number of earners in the family by 1 a decrease of -0.0087 could be seen in the job satisfaction but we already discussed that this predictor is not significant. Similarly, the regression coefficient of SATFIN shows that with a level increase in SATFIN there will be a 0.24 increase in job satisfaction. In case of WRKSLF since it has two categories (1 self employed and 2 works for somebody) if the WRKSLF changes from self employed to works for somebody then there will be -0.24 decrease in job satisfaction.

The degrees of freedom are 28670. However, the R-squared value of the model is 0.07 which is small hence our model is not a good model or it only explains 7% of the variation in job satisfaction.

Based on p-value, effect size and confidence intervals it seems that three of our hypotheses that respondent’s income, financial satisfaction and being self employed have a statistically significant effect on job satisfaction. However, as discussed in the lecture, p-values depend on both the magnitude of the relationship and the sample size. The statistical significance obtained is likely a result of our large sample and so does not necessarily mean there is a practical significance.

**[Question no. 8]**

a) Linear relationships between IVs and DV

From the plots below it seems that the relationship between the independent variables and the dependent variables is linear though the linear relationship is not very strong. For Job satisfaction and work for self the relationship seems to be negative whereas for the rest of the independent variables the relationship with the dependent variable seems to be positive.

**Code:**

# linear relationships

# take a random subset of cases (N = 100)

set.seed(3)

data.temp <- GSS.data.2[sample(1:nrow(GSS.data.2), 100),]

plot(jitter(data.temp$RINCOME, factor = 1), jitter(data.temp$JobSat, factor = 1),col="red",pch=19)

abline(lm(JobSat ~ RINCOME, data = GSS.data.2),col="blue")

plot(jitter(data.temp$EARNRS, factor = 1), jitter(data.temp$JobSat, factor = 1),col="orange",pch=19)

abline(lm(JobSat ~ EARNRS, data = GSS.data.2))

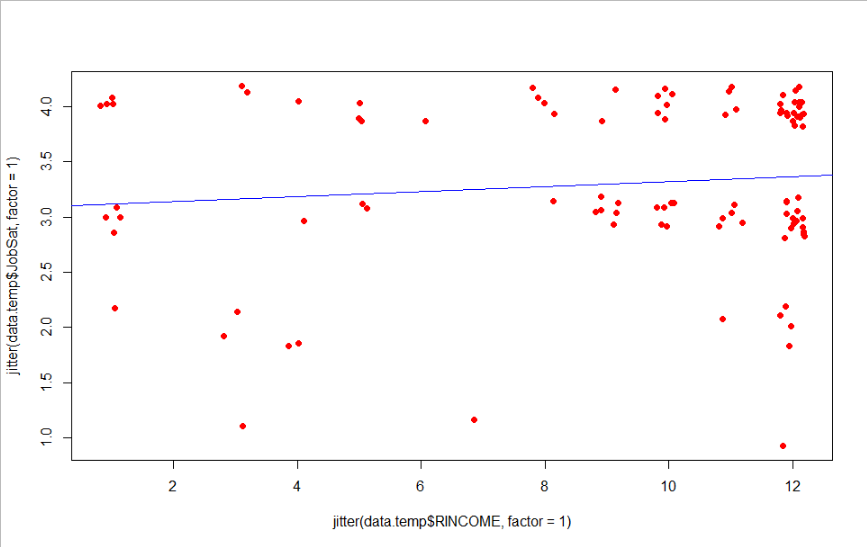
plot(jitter(data.temp$SATFIN, factor = 1), jitter(data.temp$JobSat, factor = 1),col="purple",pch=19)

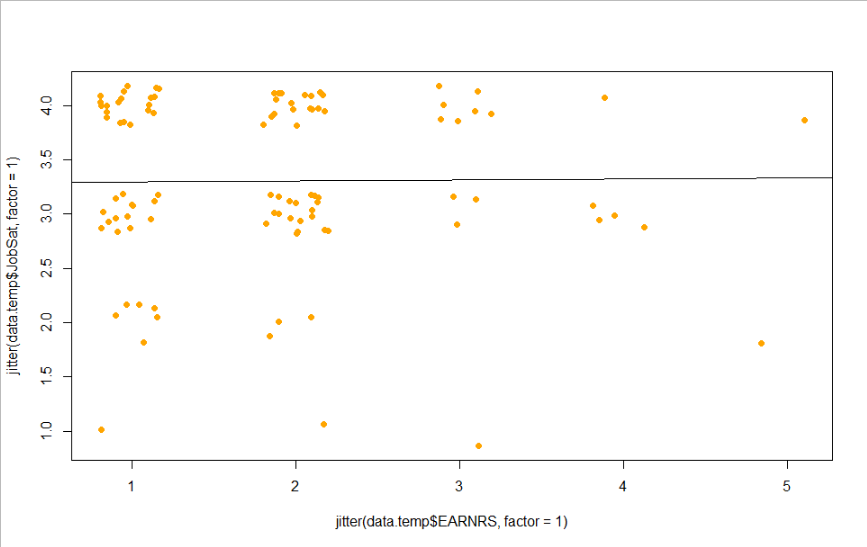
abline(lm(JobSat ~ SATFIN, data = GSS.data.2),col='red')

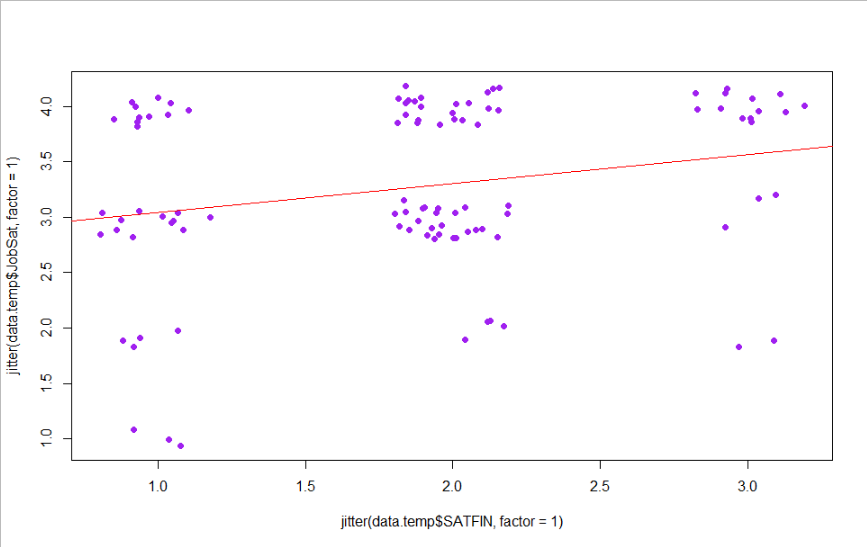
plot(jitter(data.temp$WRKSLF, factor = 1), jitter(data.temp$JobSat, factor = 1),col='green',pch=19)

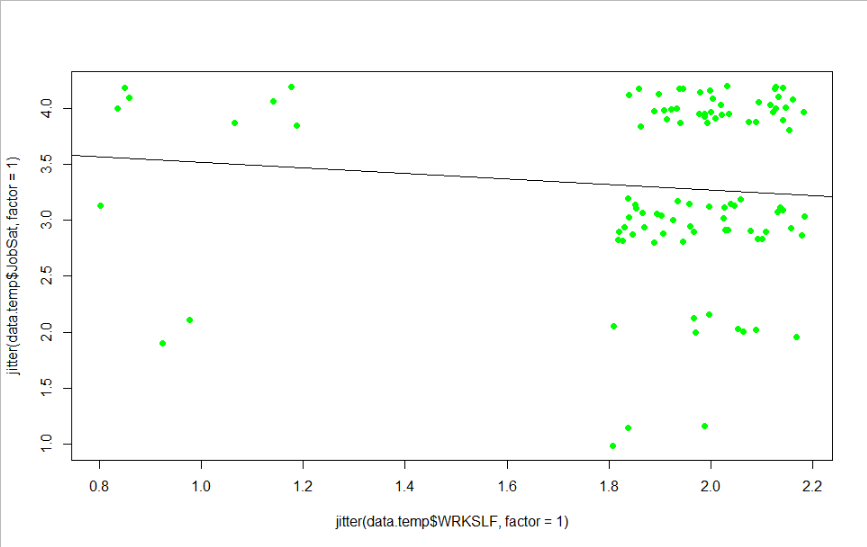
abline(lm(JobSat ~ WRKSLF, data = GSS.data.2))

**R Output:**



****

****

****

b) Multicollinearity

The variance inflation factor of all the independent variables is almost 1 hence there is no multicollinearity.

**Code:**

# multicollinearity

library(car) # this package has an easy to use multicollinearity function

vif(Model)

**Output:**

RINCOME EARNRS SATFIN WRKSLF

1.029734 1.013854 1.028455 1.002810

c) Normality of residuals

From the histogram and density plot it seems that the residuals are not completely as the plots are left skewed. Furthermore, from the normal qq plot we can see that the points do not completely follow the normal line. Hence, the assumption of normality is being violated.

**Code:**

# iid

# this works since the residuals are stored in the Model results

# but sometimes it's easier to make them a separate object we can work with by itself

Res.Model<- resid(Model)

hist(Res.Model)

# this is variation that is informative, since the histogram involves rounding

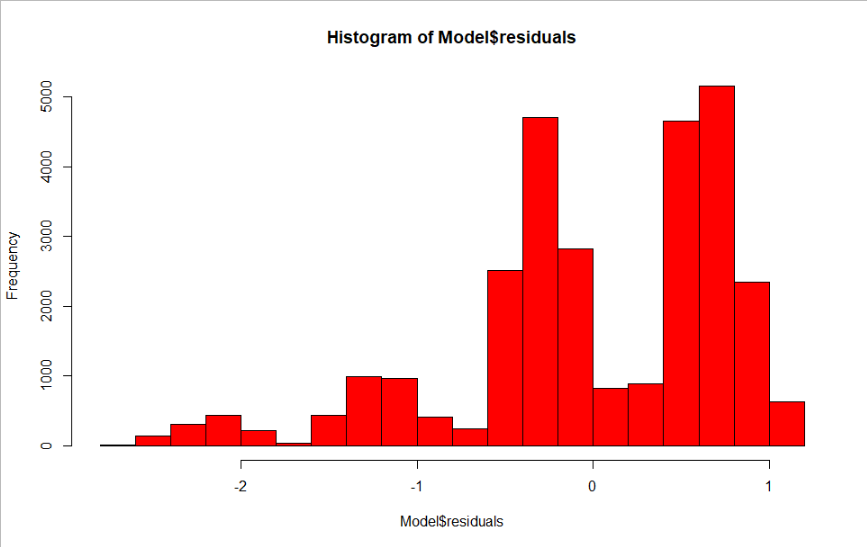
plot(density(Res.Model))

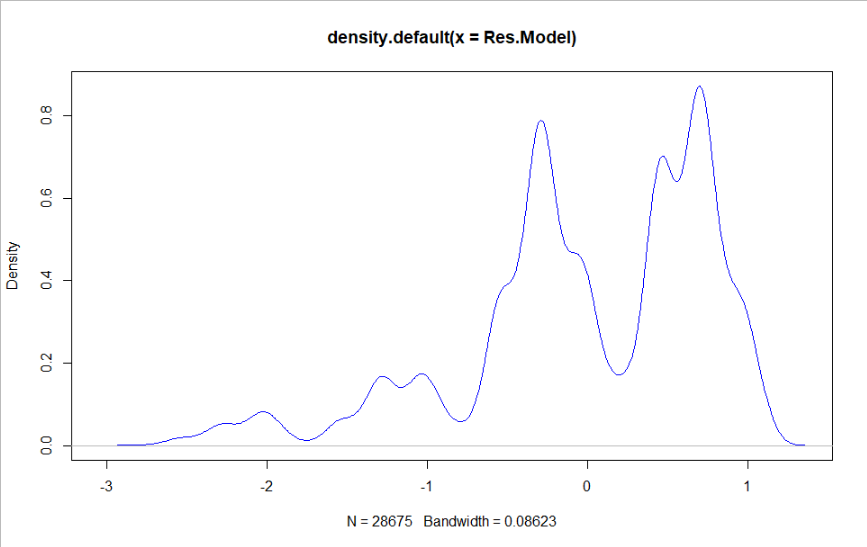
#Q-Q plot

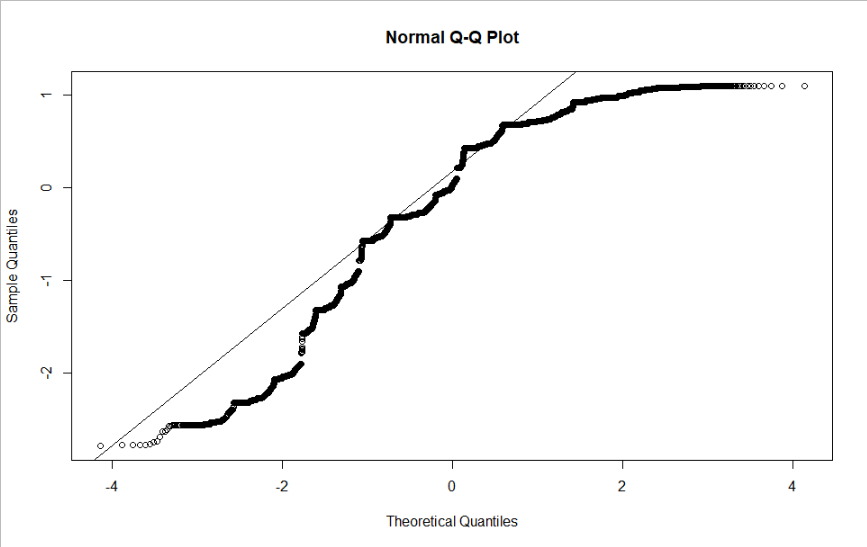
qqnorm(Res.Model) #plot the distribution against a normal curve (perfect match has all points on the line)

qqline(Res.Model)

**Output:**

****

****

****

d) Presence of heteroskedasticity

The second residuals graph is not random. When we move from left to right we can see the values from negative become positive. Since there is a pattern hence heteroskedasticity exists. One of the assumptions of applying regression is that there should be no heteroskedasticity. That assumption has been violated here.

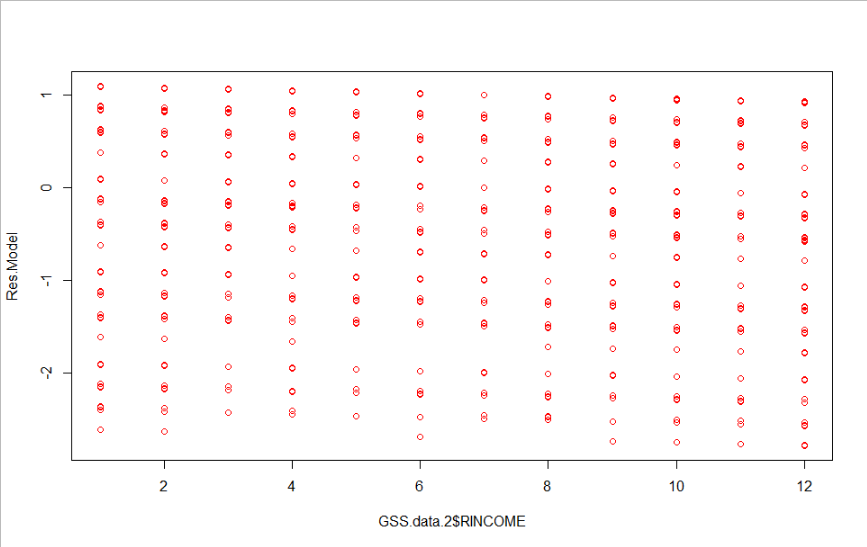
**Code:**

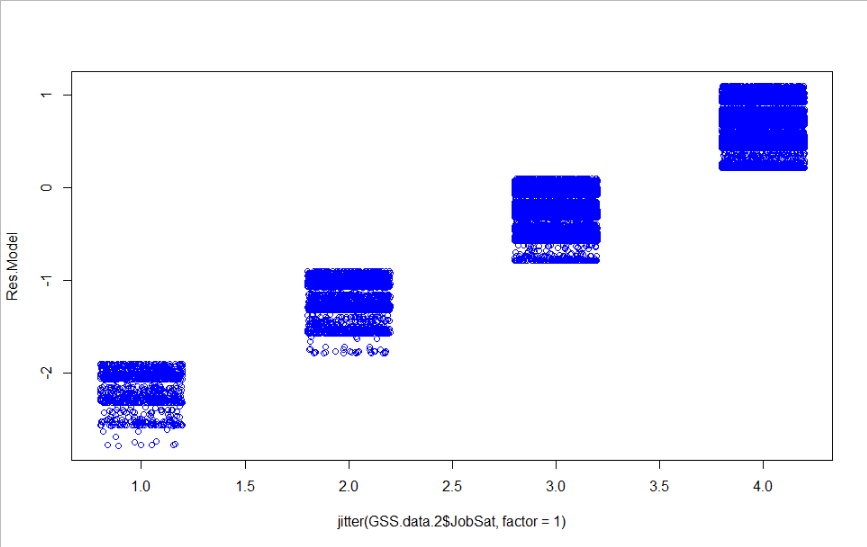
# check for heteroskedasticity

plot(GSS.data.2$RINCOME, Res.Model)

plot(jitter(GSS.data.2$JobSat, factor = 1), Res.Model)

**Output:**

****

****

e) Influential cases

Around 27231 cases are within Cook’s distance i.e. less than 4/n. 1444 points are influential points that are outside Cook’s distance. We make another model without these influential points and compare this model with the previous model. The R-squared value of the new model is 0.068 whereas the R-squared value of the previous model was 0.07. RINCOME, SATFIN and WRKSLF are still significant predictors but their effect sizes/regression coefficients have changed. The regression coefficients of RINCOME, SATFIN are still positive and the coefficient of WRKSLF is still negative in the new model but their effect sizes have decreased. The regression coefficient of EARNRS is still positive but it has slightly increased.

**Code:**

# outliers

summary(cooks.distance(Model))

# by convention, 4/n is the limit of acceptable Cook's d values

table(cooks.distance(Model) > 4/nrow(GSS.data.2))

#add the values to the data

GSS.data.2$CookD <- cooks.distance(Model)

# create a new dataset that excludes cases with high values

GSS.data.3 <- GSS.data.2[GSS.data.2$CookD <= 4/nrow(GSS.data.2),]

# compare with and without influential cases

Model.2 <- lm(JobSat ~ RINCOME + EARNRS + SATFIN + WRKSLF, data = GSS.data.3)

summary(Model)

summary(Model.2)

**Output:**

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.000e+00 3.085e-06 1.181e-05 3.480e-05 3.841e-05 1.830e-03

FALSE TRUE

27231 1444

Call:

lm(formula = JobSat ~ RINCOME + EARNRS + SATFIN + WRKSLF, data = GSS.data.2)

Residuals:

Min 1Q Median 3Q Max

-2.78587 -0.32138 0.01847 0.67862 1.09773

Coefficients:

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

(Intercept) 3.068116 0.033245 92.288 <2e-16 \*\*\*

RINCOME 0.015258 0.001378 11.069 <2e-16 \*\*\*

EARNRS 0.001482 0.005215 0.284 0.776

SATFIN 0.248307 0.006289 39.483 <2e-16 \*\*\*

WRKSLF -0.214704 0.013960 -15.380 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7692 on 28670 degrees of freedom

Multiple R-squared: 0.07012, Adjusted R-squared: 0.07

F-statistic: 540.5 on 4 and 28670 DF, p-value: < 2.2e-16

Call:

lm(formula = JobSat ~ RINCOME + EARNRS + SATFIN + WRKSLF, data = GSS.data.3)

Residuals:

Min 1Q Median 3Q Max

-2.3828 -0.3794 -0.1103 0.6172 0.8897

Coefficients:

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

(Intercept) 3.452926 0.029552 116.843 < 2e-16 \*\*\*

RINCOME 0.006302 0.001210 5.207 1.93e-07 \*\*\*

EARNRS 0.007732 0.004707 1.643 0.1

SATFIN 0.195443 0.005495 35.567 < 2e-16 \*\*\*

WRKSLF -0.276040 0.012305 -22.432 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6473 on 27226 degrees of freedom

Multiple R-squared: 0.06886, Adjusted R-squared: 0.06872

F-statistic: 503.3 on 4 and 27226 DF, p-value: < 2.2e-16

**[Question no. 9]**

Based on the above results it is clear that if we wish to continue our study we cannot progress with the current model. The current model had a low R-squared value. Furthermore, our residuals were not iid and the data was not completely normal. We will need to do some corrections for removing heteroscedasticity and also for making the data normal. Changing the way we are dealing with the data would be helpful. We can perhaps remove the number of earners variable (EARNRS) since it was not significant. We can also change the way we are dealing with the data since we were limited to 4 values in case of job satisfaction. We can perhaps increase the number of values and do the same thing with other independent variables as well. We can also change the job satisfaction variable to two values 0 and 1 to reflect whether the respondent is satisfied or not and instead of linear regression we can apply logistic regression. Our analysis would be slightly changed since we would be focusing on the aspect of whether the respondent is satisfied or not instead of how satisfied or dissatisfied he is.

In addition to modifying the selection of independent variables used, we can also explore the need for interaction terms in our analysis. We could also consider the use of feature engineering to create new variables. An example of this could be the creation of a variable ‘respondent’s income as a percentage of family income’ obtained by dividing RINCOME with INCOME.

**[Question no. 10]**

The relationship between income and job satisfaction is the subject of a vast body of research. Our results corroborate previous studies in finding that people with higher incomes are more likely to be satisfied with their job. Our findings indicate that the strength of this relationship is marginal and as discussed in the following paragraphs, a person’s satisfaction with their current financial situation is a much better indicator of job satisfaction than their income.

To our knowledge there are no previous works that have investigated the relationship between the number of earners in a household and job satisfaction. By building on the findings of Nielsen and Smyth, we expected to find that people with a higher number of earners in their household would have higher job satisfaction. However, in our analysis this variable proved to be non-significant and contrary to our expectations, had a marginal negative effect on job satisfaction. Our findings indicate that the number of earners is not an indicator of job satisfaction.

Our analysis indicates that people who are self-employed are more likely to be satisfied with their jobs. This is in-line with the findings of previous works as well as our expectations. The relatively high effect size and statistical significance of this variable indicate there is a strong relationship between being self-employed and job satisfaction. Further investigation is recommended to establish the practical significance of this relationship.

We found that there is a strong relationship between a person’s satisfaction with their financial situation and their job satisfaction. Out of the four independent variables investigated, this proved to be the best indicator of job satisfaction as indicated by its statistical significance and high effect size. This corresponds to our expectations. We recommend further investigation to confirm the practical significance of this relationship.

# **References**

Nishida, K. How to Create Table 1 Summary Information. Exploratory.io. Retrieved 13 September 2020, from https://exploratory.io/note/kanaugust/How-to-Create-Table-1-Summary-Information-YYf5BPf4Xs#:~:text=Table%201%20is%20a%20common,generate%20the%20Table%201%20information.

Cumming, G. (2012). Understanding The New Statistics.

Judge, T., Piccolo, R., Podsakoff, N., Shaw, J., & Rich, B. (2010). The relationship between pay and job satisfaction: A meta-analysis of the literature. *Journal Of Vocational Behavior*, *77*(2), 157-167. https://doi.org/10.1016/j.jvb.2010.04.002

Mathou Tchinda, J., & Bing, H. (2017). Job Satisfaction of Self-Employed Workers: Empirical Evidence from Cameroon Data. *INTERNATIONAL JOURNAL OF MANAGEMENT SCIENCE AND BUSINESS ADMINISTRATION*, *3*(6), 43-58. https://doi.org/10.18775/ijmsba.1849-5664-5419.2014.36.1006

Nielsen, I., & Smyth, R. (2008). Job satisfaction and response to incentives among China's urban workforce. *The Journal Of Socio-Economics*, *37*(5), 1921-1936. https://doi.org/10.1016/j.socec.2008.02.008