Over-Reporting in Handwashing Self-Reports: Potential Explanatory Factors and Alternative Measurements

STATS-512-Project Report

Nitasha Fazal

Introduction

Diseases like Diarrheal and respiratory infections are the main cause of fatality among young children in developing countries. Regular handwashing with soap can be an effective parameter to decrease the rate, and several programs have been initiated by health organizations to educate children and adults of developing countries on the standardized ways of handwashing.

However, measuring handwash behavior can be difficult, as direct observations (can be implemented in households and hospitals), which not only is costly and demanding but also tend to distort the natural behavior, on the other hand self-reporting which is primarily the most adopted and desirable method tend to be inflated when compared to observed data. This gives rise to over-reporting in self-reporting.

Socially desirable responding is thought to be the main source of this bias, but there can be many other sources of inflated self-report. Investigation of these socially desirable factors and factors other than social desirability is the aim of this project.

The research team analyzed the dominant predictors/factors by employing both Pearson and Point Biserial correlation methods. Their hypothesis rested on the premise that the strength of a factor's dominance increases in tandem with its higher correlation with over-reporting. They

then tested these predictors in different models for amount of variance in over-reporting explained by socially desirable responding and other factors and found that there was about 50 to 60% overreporting, also over-reporting was higher for food- than stool-related times. The highest mean value was found for handwashing before eating, while the lowest mean value was found for handwashing after defecation or urination.

The study was conducted in Ethiopia, focusing on females responsible for childcare and cooking activities. 554 primary caregivers were surveyed, however only 542 participants were observed for at least one of the foods related key times. Self-reported handwashing was measured by 0 and 1 scale (0 = never, 1 = always, for key times. Overreporting was calculated by subtracting observed scores from self-reporting scores, and the response scale range from (-1 ,0, 1) = (100% underreporting, accurate reporting, over-reporting).

We are extending this research with the aim to explore other methods and models, other than already used in the primary research, to assess whether we can develop a more effective model and refine our predictors.

Statistical Methods

Exploratory data analysis showed some missing values which were being removed (Figure 1a, b), after which we were left with 541 observations. We started our analysis by replicating the same steps and procedures being used in original research paper, that is correlation and fitting linear model. (Figure 2) The correlation results were similar but not same, primary reason of which could be different sample size and use of Point Biserial Correlation. Using Pearson correlation our results showed that for food related over reporting, Presence of

Other Adults, Group Attach, Presence of Spouse, Frequency of task Interruption were statistically significant predictors.

Initial diagnostic plots and residual plots didn't show any violation of assumptions for linear models (Figure 3). While not entirely random, the uniform variance observed in this clustering, particularly around small or large values, can be attributed to the nature of data collection through surveys, where participants are constrained by the survey scale limits, which may account for this pattern. QQ plot also shows a negative / left skew for normality of data, but its not troublesome, which means no transformation is required to proceed further. There are no influential points, which are not common for survey related data, where observations and responses are bound within limits. Polynomial transformation was also not required as assumption of linearity is not violated.

The results of linear model with same predictors used in the original research, showed that we have no evidence for most of the factors except for Descriptive and Injunctive Food Norm, and Rationalization, for which we have moderate to strong evidence (Figure 4, 6)

The value of R square suggested that about 18% variation in the response is explained by the predictors of the linear model. We also checked multicollinearity using VIF, (Figure 5) and there was no statically significant correlation among predictors.

After that, to extend the research in pursuit of refining the model, we used automated model selection technique AIC to find the optimal model among all possible models. AICs provide "evidence" or "support" for models relative to others - it does not test for differences.

Instead of adjusting interactions and additive model into same dredge, we used 2 dredge models, one to find the best predictors, next we tried the 2-way interaction among the predictors being selected from previous dredge with lowest AIC results.

Statistical Findings

Dredge function explored 256 models and the top model with lowest AIC results had all predictors included except for the MCSC (Marlow Crown Social Desirability Scale) but we selected the best model, top 2nd model as it is less complex with predictors: Injunctive, descriptive food norms, rationalization and Freq of task interruption and AIC 182.5. The difference of AIC from top model and our selected model is 0.5 units of AIC. Hence the final selected model is as follows.

$\hat{\mu}$ (Over Reporting Food | Descriptive Food Norm, Injunctive Food Norm, Freq of Task Interrupt, Rationalization) = 0.11380 + 0. 3377.Desc +0. 2665.Inj -0.09472.Freq -0.09195.Rat

Where Desc: Descriptive Norm, Inj: Injective norm, Freq: Frequency of Tas interrupt, and Rat is Rationalization

Compared to mean only model with AIC 281.8, the selected model is about 100 AIC unit less, which provides strong evidence to select this model versus the mean only model.

We conducted a comparison between our chosen model and the 2-way interaction models generated by the dredge function. Even though the AIC for the interaction models was even lower than that of our initially selected model (Figure 8), the top three 2-way interaction models, with the lowest AIC, included additional 5 interaction effects, rendering them to be overly complex. Considering the trade-off between complexity and a marginal 10-unit AIC difference

from our best model, we opted for simplicity and selected the model without any interaction terms. The AIC results are depicted in Figure 8.

Effect plots for the top selected model (Figure 9a) and interaction model (Figure 9b) also support our selected model. Effects plot show evidence that the top selected model (without interaction) has minimal violations of assumptions of linearity, equal variance, and normality, whereas for interaction model, there are few curvatures in equal variance, though we do not have strong evidence against linearity, our selected model gives better results with less computation and complexity.

Scope of Inference

Participants in the original research were not randomly selected; instead, they were selected from a specific area for the sake of accessibility and convenience. The demographics and social stratification of the participants were also very similar, which means their education level, financial conditions and social environment were similar. For this project, we are also limiting our research and analysis only to Food related overreporting.

Hence the results of the study cannot be generalized outside of this domain.

Project also doesn't deal with any random assignment; hence causal inferences cannot be deduced.

Appendix

Figure 1a (EDA and Data Cleaning)

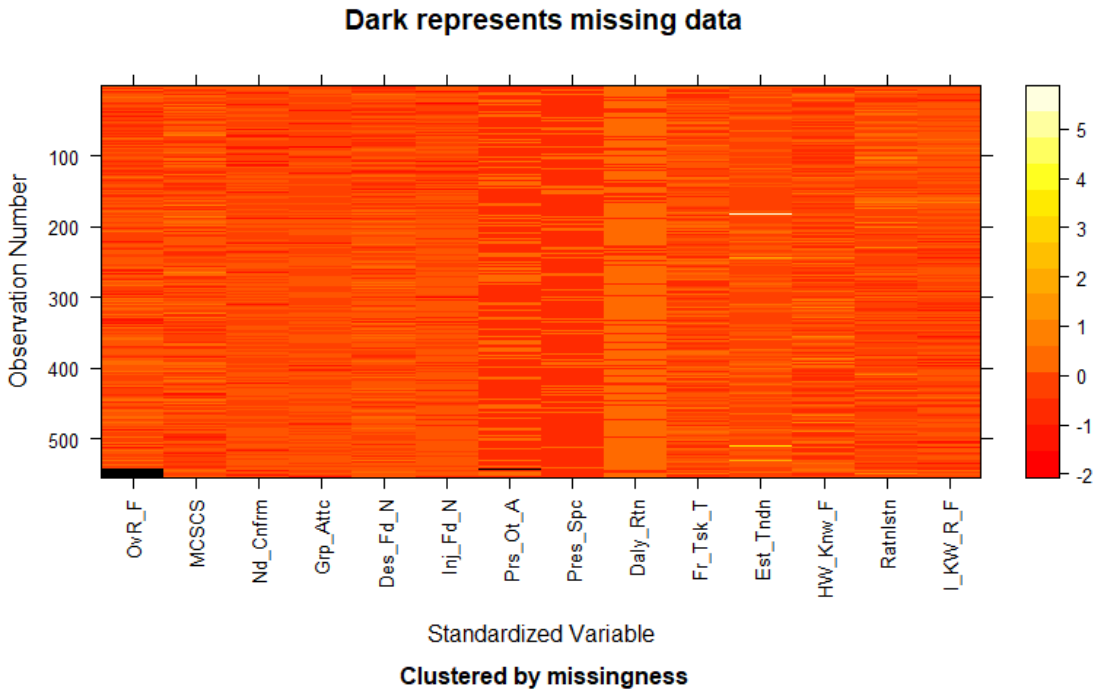


Figure 1 b

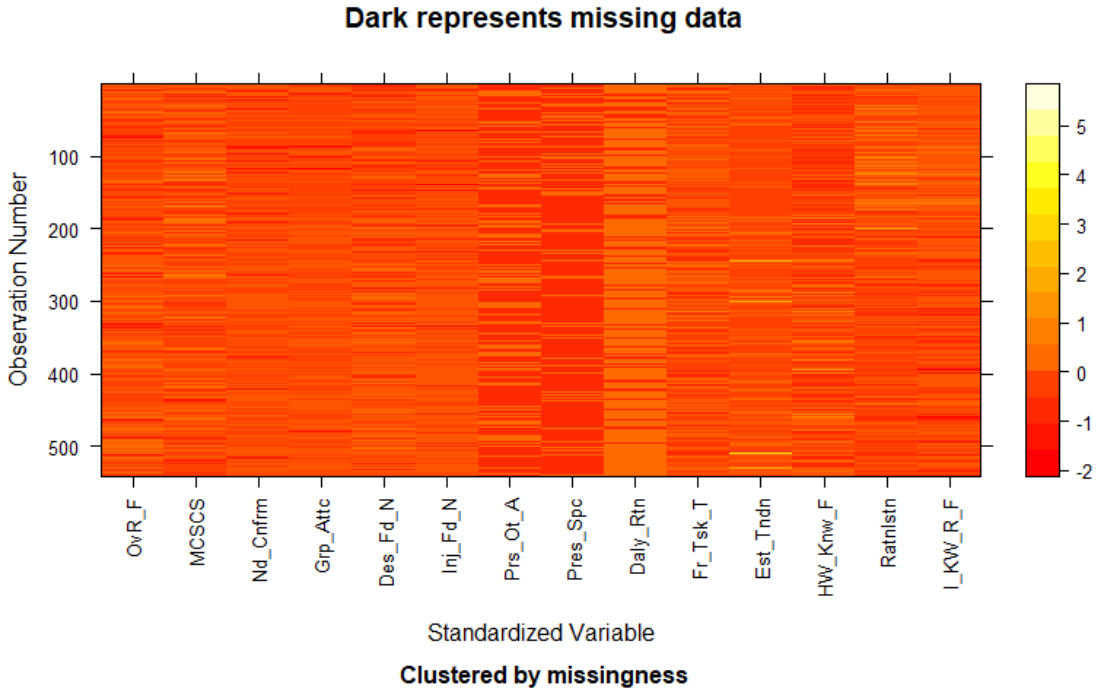


Figure 2 (Correlation among Predictors)

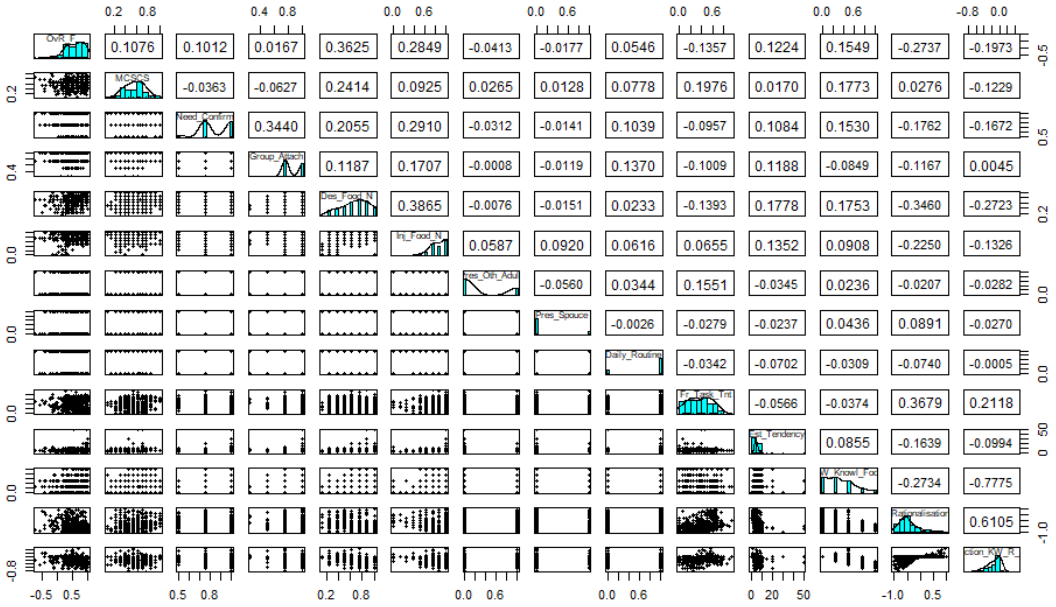

OvR_T
MCSAS
Need_Confirm
Group_Reach
Des_Food_N
Int_Food_N
Pres_Oth_Adul
Pres_Spouce
Daily_Routine
Fr_Pres_Tnt
Set_Tendency
W_Know_Foo
Rationalisation
ction_KW_R
0.1076 0.1012 0.0167 0.3625 0.2849 -0.0413 -0.0177 0.0546 -0.1357 0.1224 0.1549 -0.2737 -0.1973
-0.0363 -0.0627 0.2414 0.0925 0.0265 0.0128 0.0778 0.1976 0.0170 0.1773 0.0276 -0.1229
0.3440 0.2055 0.2910 -0.0312 -0.0141 0.1039 -0.0957 0.1084 0.1530 -0.1762 -0.1672
0.1187 0.1707 -0.0008 -0.0119 0.1370 -0.1009 0.1188 -0.0849 -0.1167 0.0045
0.3865 -0.0076 -0.0151 0.0233 -0.1393 0.1778 0.1753 -0.3460 -0.2723
0.0587 0.0920 0.0616 0.0655 0.1352 0.0908 -0.2250 -0.1326
-0.0560 0.0344 0.1551 -0.0345 0.0236 -0.0207 -0.0282
-0.0026 -0.0279 -0.0237 0.0436 0.0891 -0.0270
-0.0342 -0.0702 -0.0309 -0.0740 -0.0005
-0.0566 -0.0374 0.3679 0.2118
0.0855 -0.1639 -0.0994
-0.2734 -0.7775
0.6105

Figure 3

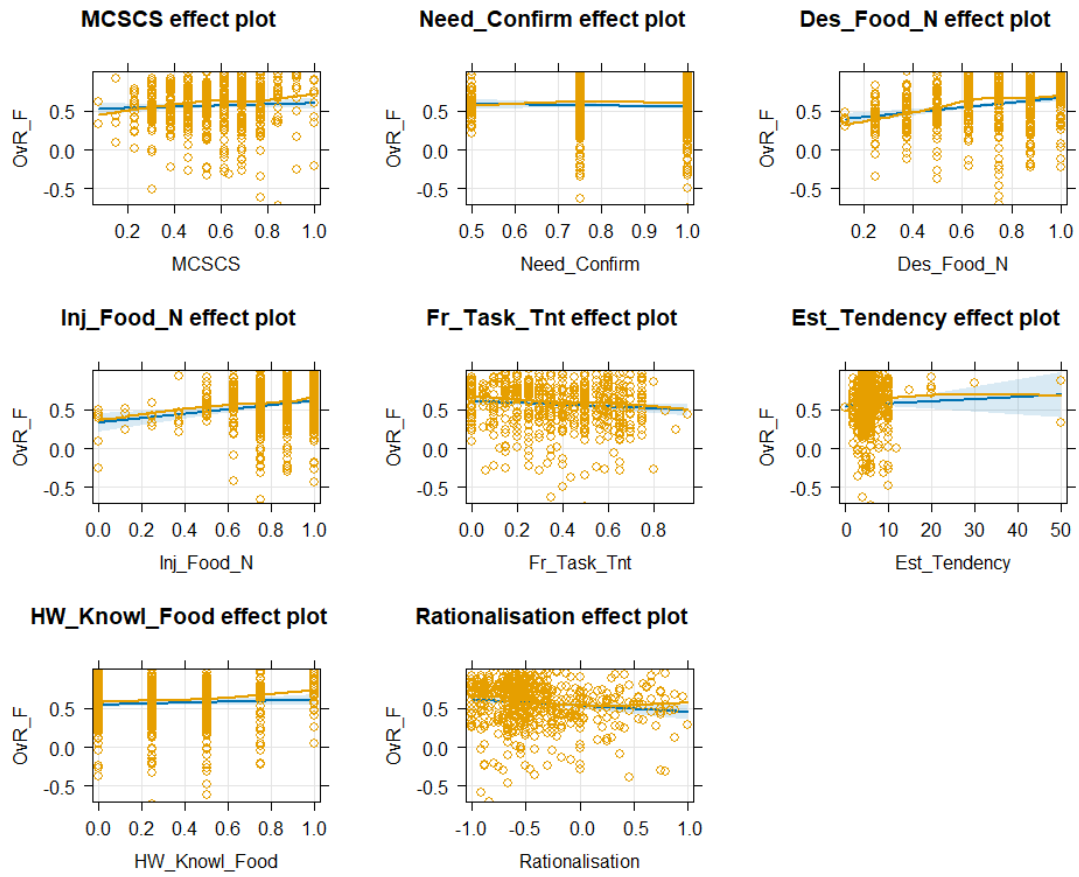


Figure 4

```
lm_F_S_R = lm(OvR_F~MCSCS + Need_Confirm  + Des_Food_N + Inj_Food_N + Fr_Task
_Tnt  +Est_Tendency+ HW_Knowl_Food + Rationalisation,  data=mydata2_b)
summary(lm_F_S_R)

##
## Call:
## lm(formula = OvR_F ~ MCSCS + Need_Confirm + Des_Food_N + Inj_Food_N +
##     Fr_Task_Tnt + Est_Tendency + HW_Knowl_Food + Rationalisation,
##     data = mydata2_b)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.29295 -0.14672  0.06111  0.19419  0.52861
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.115953   0.088697   1.307 0.191682
## MCSCS             0.077700   0.074566   1.042 0.297874
## Need_Confirm     -0.071718   0.086565  -0.828 0.407770
```

```
## Des_Food_N       0.309764   0.064010   4.839 1.71e-06
## Inj_Food_N       0.280178   0.071769   3.904 0.000107
## Fr_Task_Tnt     -0.119453   0.063399  -1.884 0.060091
## Est_Tendency     0.002779   0.003306   0.841 0.400927
## HW_Knowl_Food    0.066188   0.045586   1.452 0.147110
## Rationalisation -0.080423   0.033620  -2.392 0.017097
##
## Residual standard error: 0.2843 on 532 degrees of freedom
## Multiple R-squared:  0.1877, Adjusted R-squared:  0.1754
## F-statistic: 15.36 on 8 and 532 DF,  p-value: < 2.2e-16
```

Figure 5

```
vif(lm_F_S_R)
```

```
##           MCSCS    Need_Confirm      Des_Food_N      Inj_Food_N     Fr_Tas
k_Tnt
##        1.170463        1.144668        1.414608        1.304688        1.2
59498
##    Est_Tendency   HW_Knowl_Food Rationalisation
##        1.053225        1.142960        1.422405
```

Figure 6

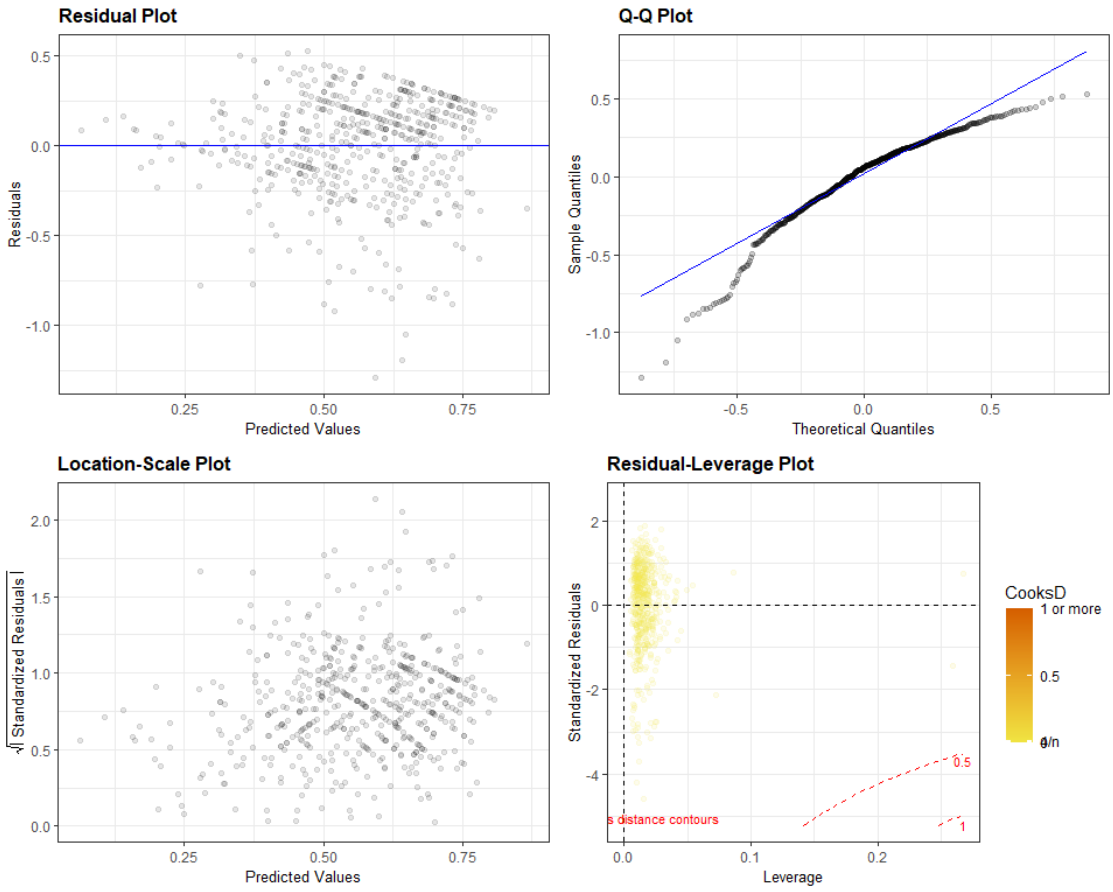


Figure 7

```
MMFood <- lm(OvR_F ~1,data=mydata2_b)
AIC(MMFood)

## [1] 281.8272

head(res_lm_F_S_R,3)

## Global model call: lm(formula = OvR_F ~ MCSCS + Need_Confirm + Des_Food_N
+ Inj_Food_N +
##     Fr_Task_Tnt + Est_Tendency + HW_Knowl_Food + Rationalisation,
##     data = mydata2_b)
## ---
## Model selection table
##       (Int) Des_Fod_N Fr_Tsk_Tnt HW_Knw_Fod Inj_Fod_N     MCS      Rtn
R^2
## 158 0.10410    0.3290   -0.10200    0.07116    0.2679         -0.07921 0.1
838
## 150 0.11380    0.3377   -0.09472               0.2665         -0.09195 0.1
799
## 182 0.07746    0.3147   -0.11300               0.2691 0.09885 -0.09362 0.1
```

```
827
##     df  logLik   AIC delta weight
## 158  7 -83.986 182.0  0.00  0.409
## 150  6 -85.270 182.5  0.57  0.308
## 182  7 -84.352 182.7  0.73  0.284
## Models ranked by AIC(x)
```

Figure 8 (Two way interaction with the previous selected model)

```
lm_F_S_R_2 = lm(OvR_F~ (Des_Food_N + Inj_Food_N + Fr_Task_Tnt  + Rationalisat
ion)^2,  data=mydata2_b)
res_lm_F_S_R_2 <- dredge(lm_F_S_R_2, rank = "AIC", extra = "R^2")
head(subset(res_lm_F_S_R_2, delta<6),3)
```

```
## Global model call: lm(formula = OvR_F ~ (Des_Food_N + Inj_Food_N + Fr_Task
_Tnt +
##     Rationalisation)^2, data = mydata2_b)
## ---
## Model selection table
##      (Int) Des_Fod_N Fr_Tsk_Tnt Inj_Fod_N     Rtn Des_Fod_N:Fr_Tsk_Tnt
## 256 0.2492    0.4021    0.06988   -0.3103 -0.2547               -1.098
## 512 0.2739    0.3620    0.03327   -0.3562 -0.3142               -1.044
## 768 0.2400    0.4986    0.03494   -0.3257 -0.1721               -1.132
##     Des_Fod_N:Inj_Fod_N Des_Fod_N:Rtn Fr_Tsk_Tnt:Inj_Fod_N Fr_Tsk_Tnt:Rtn
## 256              0.5390        0.2609               0.6749
## 512              0.5389        0.2377               0.7627          0.155
## 768              0.4625        0.3067               0.7463
##     Inj_Fod_N:Rtn    R^2 df  logLik   AIC delta weight
## 256               0.2045 10 -77.021 174.0  0.00  0.491
## 512               0.2056 11 -76.640 175.3  1.24  0.264
## 768       -0.1341 0.2054 11 -76.717 175.4  1.39  0.245
## Models ranked by AIC(x)
```

Figure 9 a

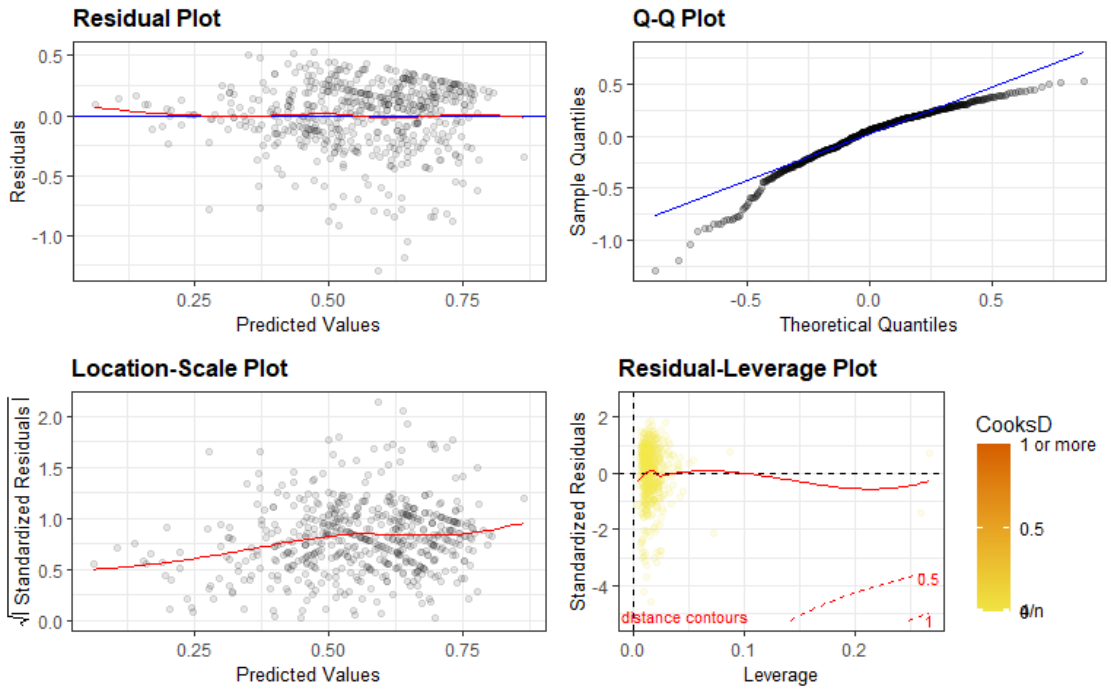
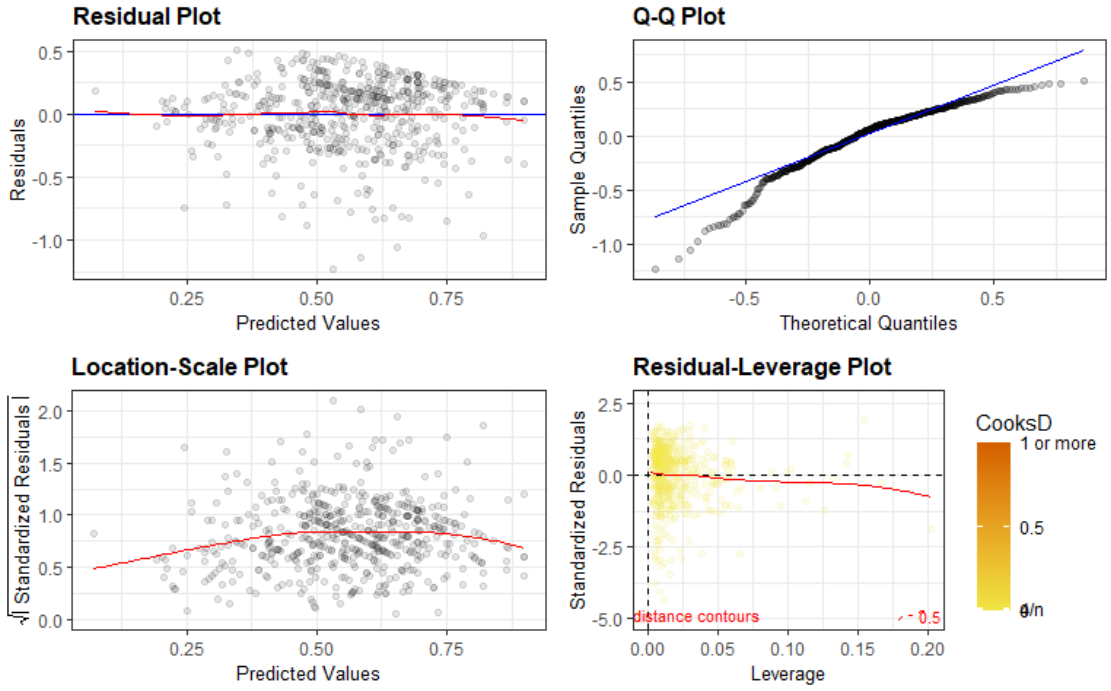


Figure 9 b



References

Research question and research paper used as base article:
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0136445
Data collected from
https://figshare.com/articles/dataset/Over_reporting_in_handwashing_self_reports_Potential_explanatory_factors_and_alternative_measurements/1304955

R-code

```
rmd<-read_lines("STAT512ProjectFinal.Rmd")
cat(paste(rmd,"\n"))

## ---
##  title: "STAT X12 Project"
##  output:
##    word_document:
##      fig_height: 5
##      fig_width: 8
##  date: ""
##  author: Name
##  ---
##
##  ```{r setup, include=FALSE}
##  knitr::opts_chunk$set(message = FALSE,
##                        warning = FALSE)
##  options(show.signif.stars = FALSE)
##
##  library(ggplot2)
##  library(ggthemes)
##  library(tidyverse)
##  library(car)
##  library(effects)
##  library(janitor)
##  library(readxl)
##  library(catstats2)
##  library(mosaic)
##  theme_set(theme_bw()) #Prevents need for + theme_bw() in ggplots
##  ```
##
##
##  ```{r}
##  library(mi)
##  library(haven)
##  mydata <- read_sav("C:\\Users\\nitas\\Downloads\\Overreportinginhandwashi
ngselfreports_Final.sav")
##  ```
##
```

```
##  ```{r}
##  my_data_m <- missing_data.frame(data.frame(mydata))
##  image(my_data_m)
##  colnames(my_data_m)
##  dim(my_data_m)
##  ```
##
##
##  ```{r}
##  colSums(!is.na(mydata))
##  ```
##  ```{r}
##  colnames(mydata)[which(names(mydata) == "OR_SR_food")] <- "OvR_F"
##  colnames(mydata)[which(names(mydata) == "OR_SR_stool")] <- "OvR_S"
##  colnames(mydata)[which(names(mydata) == "E315_327_SD")] <- "MCSCS"
##  colnames(mydata)[which(names(mydata) == "E304_01")] <- "Need_Confirm"
##  colnames(mydata)[which(names(mydata) == "E305_01")] <- "Group_Attach"
##  colnames(mydata)[which(names(mydata) == "E6134_01")] <- "Des_Food_N"
##  colnames(mydata)[which(names(mydata) == "E6156_01")] <- "Inj_Food_N"
##  colnames(mydata)[which(names(mydata) == "E5123_01")] <- "Des_Stool_N"
##  colnames(mydata)[which(names(mydata) == "E5145_01")] <- "Inj_Stool_N"
##  colnames(mydata)[which(names(mydata) == "E121.1_exHusb_cat")] <- "Pres_Ot
h_Adult"
##  colnames(mydata)[which(names(mydata) == "E341.1")] <- "Pres_Spouce"
##  colnames(mydata)[which(names(mydata) == "E301")] <- "Daily_Routine"
##  colnames(mydata)[which(names(mydata) == "E335_339_2_MT_01")] <- "Fr_Task_
Tnt"
##  colnames(mydata)[which(names(mydata) == "E342")] <- "Est_Tendency"
##  colnames(mydata)[which(names(mydata) == "E713_HW_stool_01")] <- "HW_Knowl
_Stool"
##  colnames(mydata)[which(names(mydata) == "E713_HW_food_01")] <- "HW_Knowl_
Food"
##  colnames(mydata)[which(names(mydata) == "E328_E334_DISS_Rationalisation_0
1")] <- "Rationalisation"
##  colnames(mydata)[which(names(mydata) == "E713_stoolxE32834")] <- "Interac
tion_KW_R_Stool"
##  colnames(mydata)[which(names(mydata) == "E713_foodxE32834")] <- "Interact
ion_KW_R_Food"
##  colnames(mydata)[which(names(mydata) == "E306_repoled_01")] <- "Loaded_Qu
e_Word"
##  colnames(mydata)[which(names(mydata) == "MeanHW_01")] <- "SR_Food_Stool"
##  colnames(mydata)[which(names(mydata) == "Version_numerisch")] <- "Sc_Que_
Version"
##  colnames(mydata)[which(names(mydata) == "E306_E307_FW_umgepolt_01")] <- "
Sc_Forgiving"
##  colnames(mydata)[which(names(mydata) == "E309_E311_E306_context_01_repol"
)] <- "Sc_Que_Context"
##  colnames(mydata)[which(names(mydata) == "OR_ICR_stool")] <- "OR_Script_St
ool"
##  colnames(mydata)[which(names(mydata) == "OR_ICR_food")] <- "OR_Script_Foo
```

```
d"
##
##
##  ```
##
##
##  Subset for food related Data
##  ```{r}
##  mydata2 <- mydata[ , c(35:ncol(mydata))]
##  mydata2 <- mydata2[ , c(-5,-7,-14,-17,-24)]
##
##
##  my_data2_m <- missing_data.frame(data.frame(mydata2))
##  image(my_data2_m)
##  colnames(my_data2_m)
##  dim(my_data2_m)
##  ```
##  Removing the script related variables from Food related dataset.
##  ```{r}
##  mydata2_a <- mydata2[ , c(-15,-16,-17,-18,-19,-20)]
##
##
##  my_data2_m <- missing_data.frame(data.frame(mydata2_a))
##  image(my_data2_m)
##  dim(my_data2_m)
##
##  mydata2_b <- mydata2_a %>% drop_na()
##  my_data2b_m <- missing_data.frame(data.frame(mydata2_b))
##  image(my_data2b_m)
##  dim(my_data2b_m)
##  ```
##
##  Correlation to find the answers of first 9 hypothesis.(Food)
##  ```{r fig.height=6,fig.width=10}
##  library(psych)
##  pairs.panels(mydata2_b,smooth=FALSE,ellipses=FALSE,digits=4)
##  {pairs.panels}
##  ```
##
##  Fitting the Linear Models
##  ```{r}
##
##  lm_F_S_R = lm(OvR_F~MCSCS + Need_Confirm  + Des_Food_N + Inj_Food_N + Fr_
Task_Tnt  +Est_Tendency+ HW_Knowl_Food + Rationalisation,  data=mydata2_b)
##  vif(lm_F_S_R)
##  summary(lm_F_S_R)
##
##  ```
##
##
```

```
##  Checking the Assumptions for the linear Model
##  ```{r fig.height=8,fig.width=10}
##
##  library(ggResidpanel)
##  library(haven)
##  mydata2_b <- zap_formats(zap_labels(mydata2_b))
##  plot(allEffects(lm_F_S_R, residuals = T), grid = T)
##  resid_xpanel(lm_F_S_R, yvar = "response")
##  resid_panel(lm_F_S_R, "R", alpha=0.1)
##
##  ```
##
##  From here we will do our own analysis on the
##
##  Research Question:
##  We will systematically investigate various model combinations to identify
potential predictors that may offer improved insights into the over reporting
of hand washing behavior."
##
##  From here on we will only work for the Food related over reporting of han
d wash.
##  Lets pick the full model according to paper.
##
##  ```{r}
##  options(na.action = "na.fail")
##  library(MuMIn) #Load the multi-model inference package
##  # prevent fitting sub-models to different datasets
##  options(na.action = "na.fail") #Must be run to use dredge
##  res_lm_F_S_R <- dredge(lm_F_S_R, rank = "AIC", extra = "R^2")
##  subset(res_lm_F_S_R, delta<6)
##  dim(res_lm_F_S_R)
##
##  ```
##
##
##  ```{r}
##
##  MMFood <- lm(OvR_F ~1,data=mydata2_b)
##  AIC(MMFood)
##  head(res_lm_F_S_R,3)
##  ```
##  There were 256 models explored. The top selected model and mean only mode
l has AIC difference of 99.82. which provides us strong evidence for our top
model.
##  Top model with AIC 182, contains Des_Fod_N , Fr_Tsk_Tnt ,HW_Knw_Fod ,Inj_
Fod_N and rationalization as predictors
##
##
##  Next we tested interaction
##  ```{r fig.height=8,fig.width=10}
```

```
##
##  lm_F_S_R_2 = lm(OvR_F~ (Des_Food_N + Inj_Food_N + Fr_Task_Tnt  + Rational
isation)^2,  data=mydata2_b)
##  res_lm_F_S_R_2 <- dredge(lm_F_S_R_2, rank = "AIC", extra = "R^2")
##  head(subset(res_lm_F_S_R_2, delta<6),3)
##  dim(res_lm_F_S_R_2)
##  ```
##
##
##  Effect Plot
##  ```{r}
##  resid_panel(model = lm_F_S_R, plot= 'R', alpha=0.1, smoother= T)
##  resid_panel(model = lm_F_S_R_2, plot= 'R', alpha=0.1, smoother= T)
##  ```
##  ```{r catRmd,asis=TRUE}
##  rmd<-read_lines("STAT512ProjectFinal.Rmd")
##  cat(paste(rmd,"\n"))
##  ```
```