Analyzing Sustainable Livelihoods in Bangladesh for World Renew

STAT-245, Calvin University

Ivanna Rodriguez, Myungha Kim, James Eapen

December 18, 2019

Contents

Data Preparation

Executive Summary

The goal of the project was to perform a comparative analysis between the Bangladesh Integrated Household Income Survey (BIHS) conducted by the International Food Policy Research Institute (IFPRI), and the GAC Livelihoods survey conducted by our client. The Sustainable Livelihoods project from which the GAC survey originates is a program that World Renew implemented in Bangladesh. Per our client's request, we only performed analysis on the BIHS dataset. The goal of this approach was to draw conclusions from this general household survey that would allow our client to compare to the conclusions they reached when evaluating their own program. Ultimately, our client was interested in measuring whether World Renew's program had an impact in Bangladesh by treating the BIHS Survey we used for this analysis as their "control" group. We found that each household's number of cattle, household total occupancy, precipitation, the ownership of a cartplough, and telephone ownership were significant at at least a 0.05 significance level in influencing the food security score. Our attempts at descriptive statistics did not work because there were too many NA values in the data that prevented us from getting representative measures from the dataset.

Our analysis is divided in two parts. First, we created a linear model that would allow us to assess whether assets are good predictors of food security. Second, we created descriptive statistics. These serve the purpose of allowing our client to compare against their results for the GAC Livelihoods program evaluation.

Questions and Answers

Were there broader changes taking place in the communities where World Renew's partners worked that influenced our observed results? We planned on answering this question using the results from World Renew and the 2011 and 2015 data sets from the Bangladesh Integrated Household Survey. However, as we made our models, we realized that we could not make good comparisons because the variables used by the GAC livelihoods regressions were quite different from the variables available to us in the BIHS dataset. We attempted to make similar variables, but were not able to because there were no variables we could use or the data had too many NA values.

Food Consumption Model

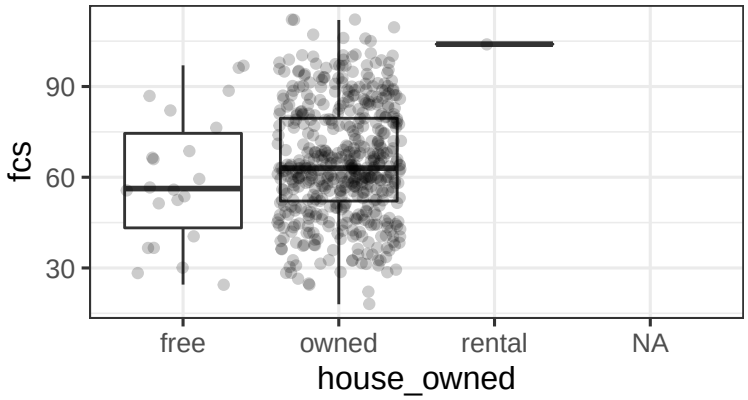
The following document outlines our process looking at the factors that are associated with variations in the Food Consumption Score (`fcs`) which is a numeric score with values ranging from 0 to 120. This score is calculated by International Food Policy Research Institute. BIHS dataset was filtered by specific regions in Bangladesh where World Renew ran their programs.

We used the variable `fcs` as our response variable because according to our client, this variable was the closest measure to the food security scores used as the response variable in the GAC Livelihoods regressions carried out by our client. To chose our predictors, we looked at the predictors deemed "statistically significant" after the hypothesis tests performed by our client, and tried to find similar predictors in the BIHS dataset. We determined that `house_owned` and all of the assets predictors would be good candidates. In addition to these variables, we decided to include demographic variables such as total number of household members and a couple biophysical predictors like percipitation (`bio_bio_12`) and temperature (`bio_bio_1`).

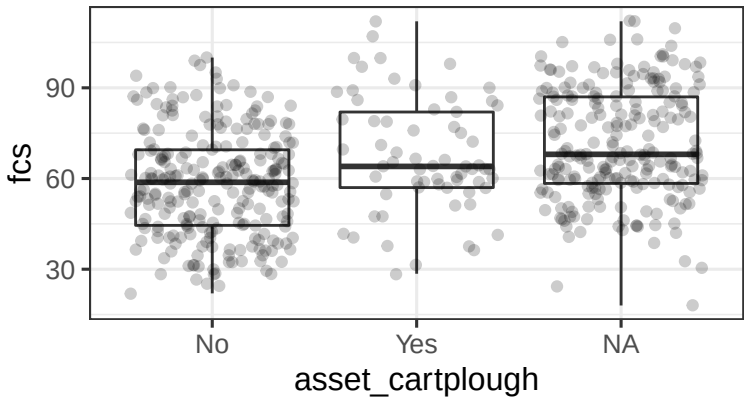
Exploratory plots

Of the variables selected for our regression, we decided to plot those that were categorical using boxplots to see if there might be a relationship between the response variable `fcs` and these predictors. We also added jitter plots to get a better idea of the sample size. The resulting plots are shown below.

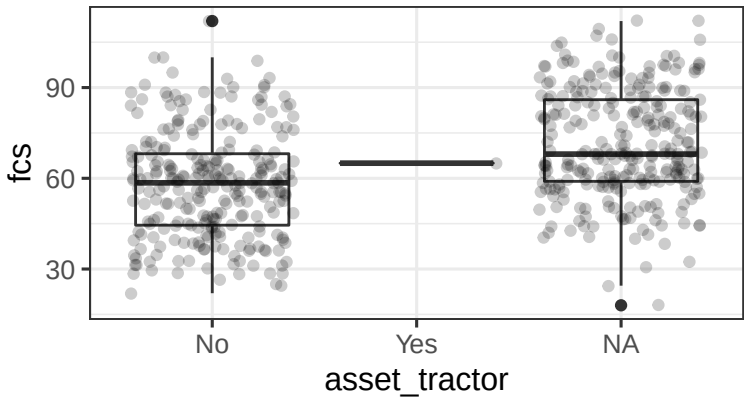
```
# whether they owned/rented a house
gf_boxplot(data = bihs, fcs ~ house_owned)%>%gf_jitter(alpha = 0.2)
```



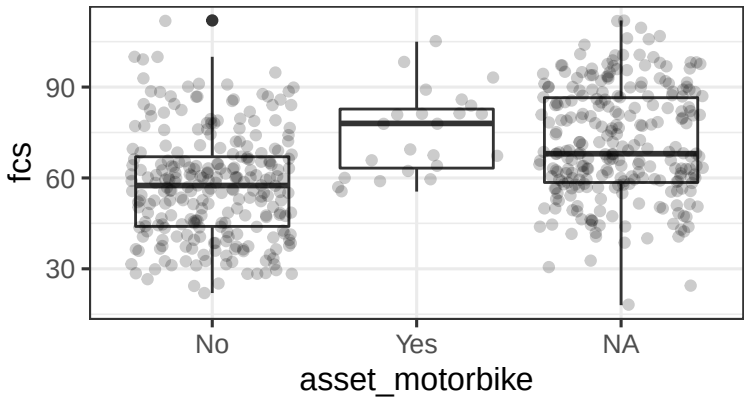

```
# tech assets
gf_boxplot(data = bihs, fcs ~ asset_cartplough)%>%gf_jitter(alpha = 0.2)
```



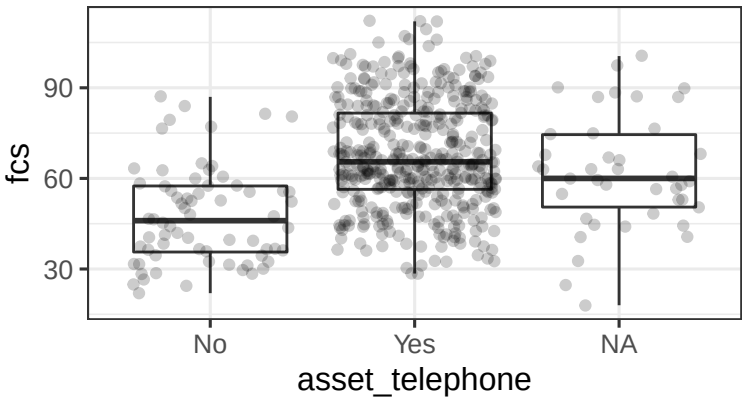

```
gf_boxplot(data = bihs, fcs ~ asset_tractor)%>%gf_jitter(alpha = 0.2)
```



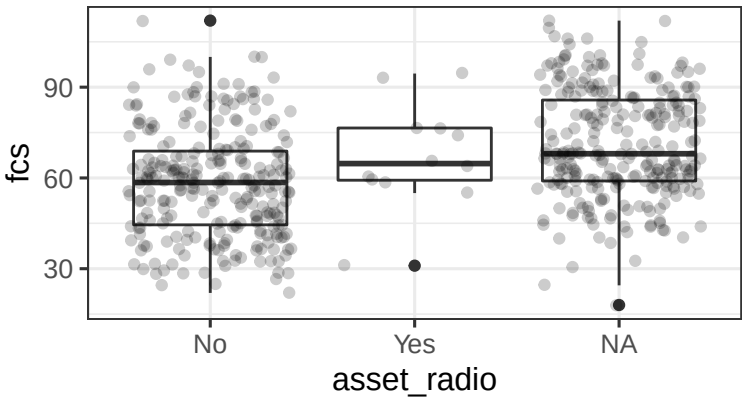

```
gf_boxplot(data = bihs, fcs ~ asset_motorbike)%>%gf_jitter(alpha = 0.2)
```




```
gf_boxplot(data = bihs, fcs ~ asset_telephone)%>%gf_jitter(alpha = 0.2)
```




```
gf_boxplot(data = bihs, fcs ~ asset_radio)%>%gf_jitter(alpha = 0.2)
```



From the boxplots above, there is a positive correltation between `fcs` and all our technological assets. In addition, for our `house_owned` variable, the plot shows higher `fcs` scores for those who own and/or rent their house. We also can see there are a lot of missing values in our dataset which is concerning.

Fitting the linear model

We encountered problems when including most of the variables into a single model. The model below is the one that worked the best. It includes our two main variables of interest: `asset_cartplough` and `house_owned`.

```
asset_lm <- lm(data = bihs_original, fcs ~ factor(survey_year) + asset_qty_poultry +
                 asset_qty_cattle + asset_qty_otherlivestock + asset_qty_sheepgoat +
                 memb_total + memb_und15 + memb_15_44 + hhs_total + bio_bio_1 +
                 bio_bio_12 + house_owned + asset_cartplough + asset_telephone, na.action = 'na.fail')

summary(asset_lm)
```

```
##
## Call:
## lm(formula = fcs ~ factor(survey_year) + asset_qty_poultry +
##     asset_qty_cattle + asset_qty_otherlivestock + asset_qty_sheepgoat +
##     memb_total + memb_und15 + memb_15_44 + hhs_total + bio_bio_1 +
##     bio_bio_12 + house_owned + asset_cartplough + asset_telephone,
##     data = bihs_original, na.action = "na.fail")
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -31.560 -10.627  -0.697   9.491  44.950
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               132.268232 353.041500   0.375 0.708197
## factor(survey_year)2015     2.824916   4.567763   0.618 0.536775
## asset_qty_poultry           0.002366   0.240925   0.010 0.992170
## asset_qty_cattle            2.758238   1.300787   2.120 0.034833 *
## asset_qty_otherlivestock   -0.365057   0.339079  -1.077 0.282563
## asset_qty_sheepgoat        -1.411200   2.072145  -0.681 0.496403
## memb_total                  2.758834   1.104307   2.498 0.013044 *
## memb_und15                 -1.840420   1.175480  -1.566 0.118534
## memb_15_44                 -1.158268   1.261428  -0.918 0.359281
## hhs_total                   4.812035   1.552080   3.100 0.002126 **
## bio_bio_1                  -3.415063  13.176284  -0.259 0.795681
## bio_bio_12                 -0.010803   0.008462  -1.277 0.202776
## house_ownedowned           -4.103573   4.314379  -0.951 0.342340
## asset_cartploughYes         4.990623   2.556769   1.952 0.051926 .
## asset_telephoneYes          8.504240   2.513287   3.384 0.000815 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.56 on 285 degrees of freedom
## Multiple R-squared:  0.2986, Adjusted R-squared:  0.2641
## F-statistic: 8.665 on 14 and 285 DF,  p-value: 1.192e-15
```

We can see that this model shows a low Adjusted R-squared value which is 0.2641. This implies that our model explains only 26 percent of the variation in the response. This is not surprising, however, as we are using a small numeber of predictors, and the BIHS survey contained more variables that were not used for the purpose of this analysis.

Checking for multicollinearity between predictors

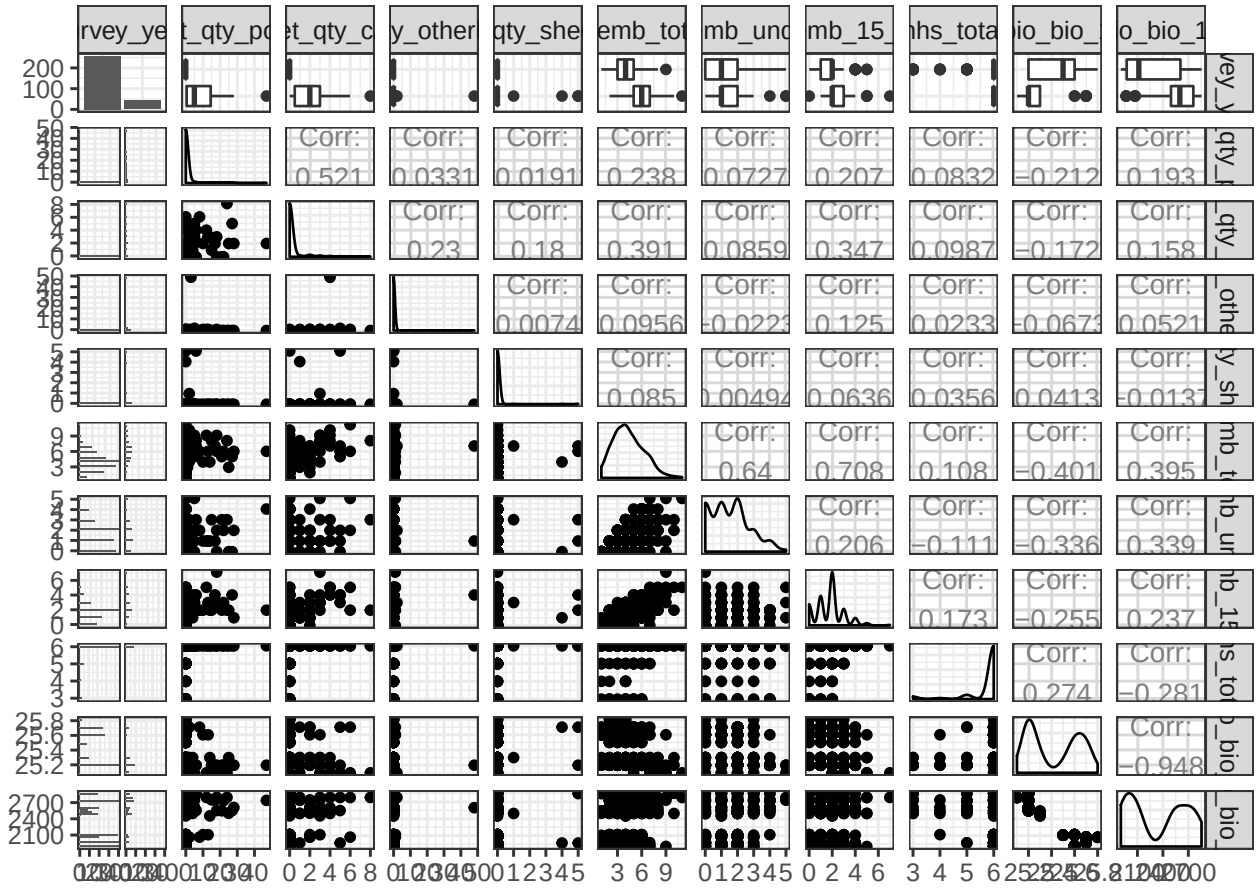
We wanted to check for multicollinearity between our predictors to eliminate the possibility of inflated variance by getting rid of highly correlated predictors.

```
# checking for collinearity in quantitative predictors
vif(asset_lm)
```

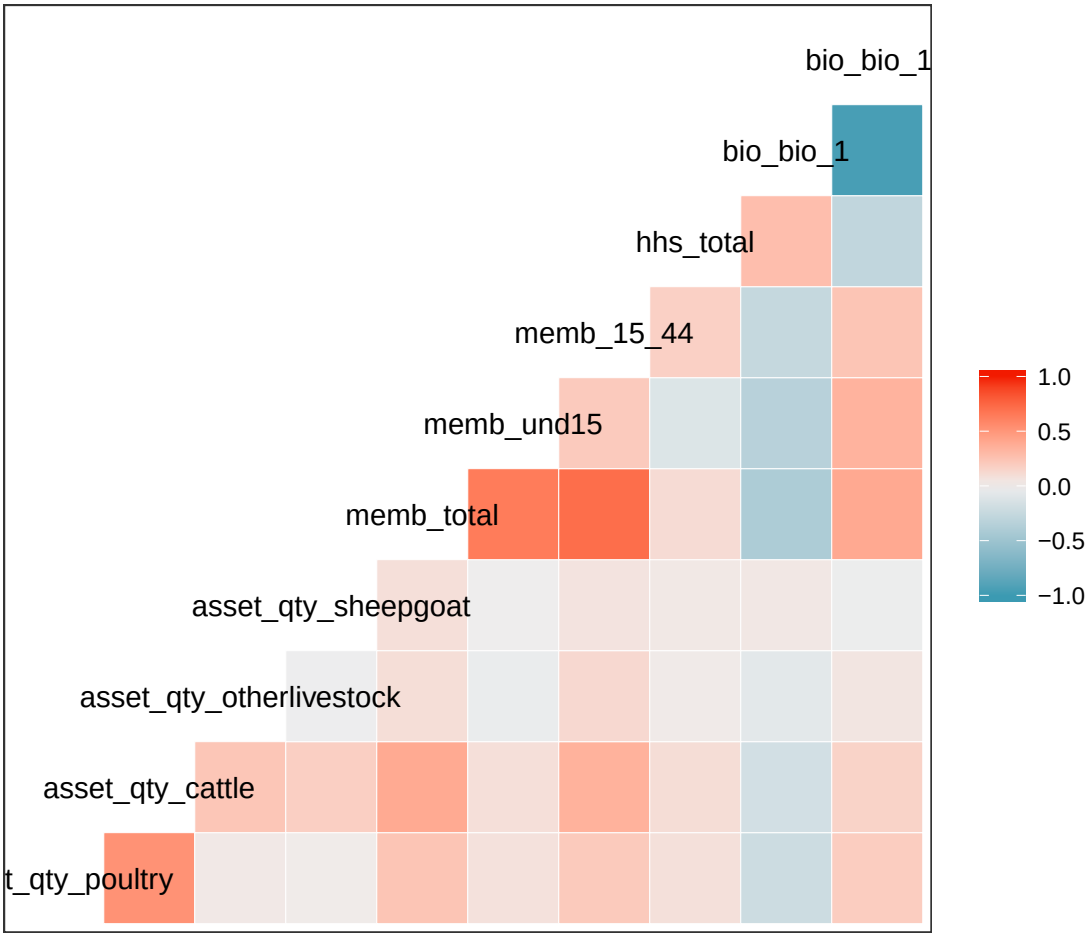
```
##      factor(survey_year)        asset_qty_poultry         asset_qty_cattle
##                 3.175399                 1.785341                 2.411222
## asset_qty_otherlivestock      asset_qty_sheepgoat               memb_total
##                 1.094819                 1.175269                 5.405350
##               memb_und15               memb_15_44                hhs_total
##                 2.550880                 2.752706                 1.334021
##                bio_bio_1               bio_bio_12              house_owned
##                10.335599                10.288256                 1.095878
##         asset_cartplough          asset_telephone
##                 1.312503                 1.372369
```

```
predictors <- bihs_original %>% select(
                              survey_year,
                              asset_qty_poultry,
                              asset_qty_cattle,
                              asset_qty_otherlivestock,
```

```
                        asset_qty_sheepgoat,
                        memb_total,
                        memb_und15,
                        memb_15_44,
                        hhs_total,
                        bio_bio_1,
                        bio_bio_12
                        )
ggpairs(predictors)
```




```
ggcorr(predictors)
```



After checking for multicollinearity, we decided to eliminate `memb_total` and `bio_bio_1` which are shown to be highly correlated both in the reported GVIF scores (scores higher than 2 are highly correlated), and the correlation matrix plots.

Fitting a new model

We decided to fit a new model without highly correlated predictors below.

```
asset_lm_two <- lm(data = bihs_original, fcs ~ factor(survey_year) + asset_qty_poultry +
                     asset_qty_cattle + asset_qty_otherlivestock + asset_qty_sheepgoat +
                     memb_15_44 + hhs_total + bio_bio_12 + house_owned + asset_cartplough +
                     asset_telephone, na.action = 'na.fail')

summary(asset_lm_two)
```

```
##
## Call:
## lm(formula = fcs ~ factor(survey_year) + asset_qty_poultry +
##     asset_qty_cattle + asset_qty_otherlivestock + asset_qty_sheepgoat +
##     memb_15_44 + hhs_total + bio_bio_12 + house_owned + asset_cartplough +
##     asset_telephone, data = bihs_original, na.action = "na.fail")
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -33.094 -10.219  -0.942   9.936  47.553
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               37.925278  14.041200   2.701 0.007323 **
## factor(survey_year)2015    5.804535   4.384948   1.324 0.186639
## asset_qty_poultry         -0.080570   0.238767  -0.337 0.736030
## asset_qty_cattle           2.961549   1.303775   2.272 0.023853 *
## asset_qty_otherlivestock -0.374860   0.340244  -1.102 0.271495
## asset_qty_sheepgoat      -1.657270   2.074249  -0.799 0.424964
## memb_15_44                 1.091252   0.907962   1.202 0.230401
## hhs_total                  5.177486   1.549232   3.342 0.000942 ***
## bio_bio_12                -0.007133   0.003233  -2.207 0.028131 *
## house_ownedowned          -2.695769   4.300807  -0.627 0.531283
## asset_cartploughYes        5.428200   2.564547   2.117 0.035149 *
## asset_telephoneYes         8.980438   2.501773   3.590 0.000389 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.65 on 288 degrees of freedom
## Multiple R-squared:  0.2825, Adjusted R-squared:  0.2551
## F-statistic: 10.31 on 11 and 288 DF,  p-value: 6.842e-16
```

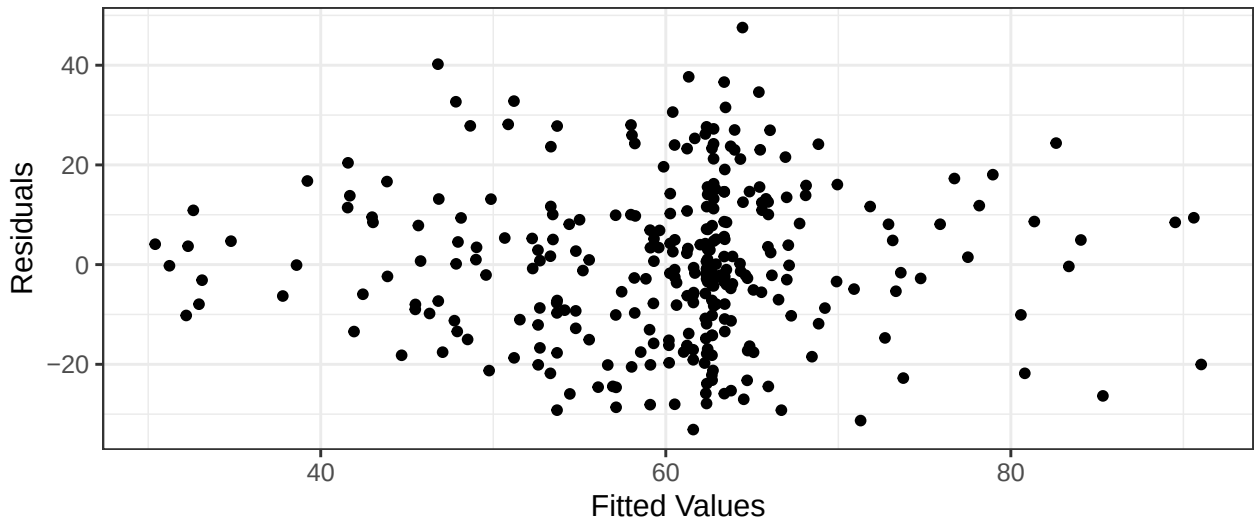
```
vif(asset_lm_two)
```

```
##      factor(survey_year)        asset_qty_poultry         asset_qty_cattle
##                 2.890833                 1.732250                 2.392948
## asset_qty_otherlivestock      asset_qty_sheepgoat               memb_15_44
##                 1.088992                 1.163380                 1.408877
##                hhs_total              bio_bio_12             house_owned
##                 1.313017                 1.483134                 1.075793
##         asset_cartplough          asset_telephone
##                 1.304493                 1.343338
```

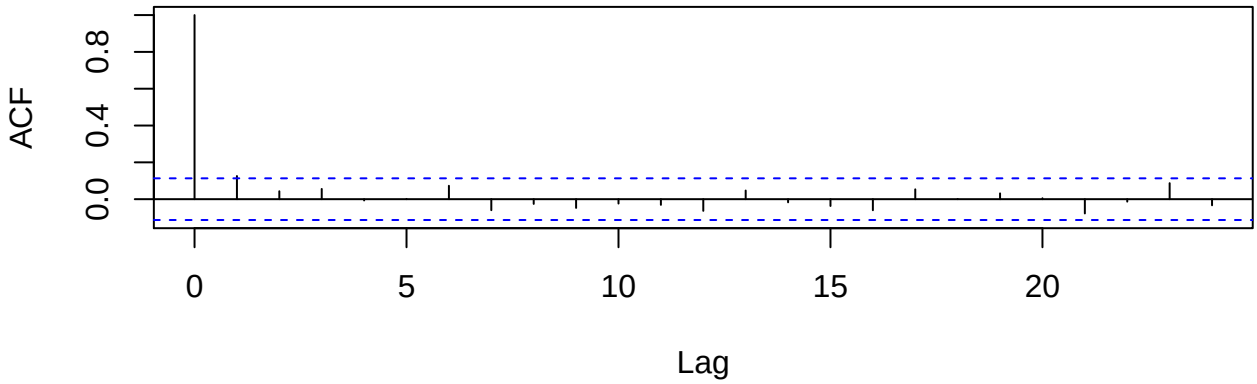
We were able to reduce the correlation problem with this model, as the GVIF scores are lower. We will use this model for the rest of this analysis.

Model assessment

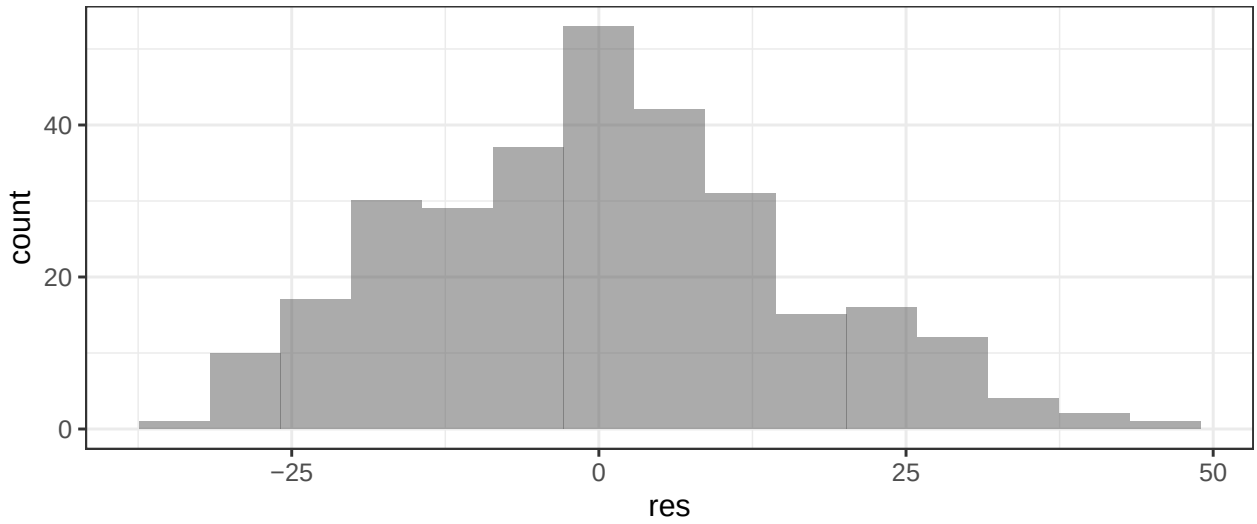
```
# linearity and constant variance
bihs_original <- bihs_original %>%
  mutate(res = resid(asset_lm_two),
         fitted = predict(asset_lm_two))
gf_point(res ~ fitted, data = bihs_original) %>%
  gf_labs(x = 'Fitted Values', y = 'Residuals')
```




```
# independence of residuals
acf(resid(asset_lm_two), main = '')
```




```
#normality of residuals
gf_histogram(~res, data = bihs_original, bins = 15) # they look a bit right skewed...
```



All the conditions for a linear regression seem to be met by our model, no major problems with linearity, constant variance, independence of residuals or normality of residuals can be seen.

Model selection

We are demonstrating which predictors are the best at explaining the response variable in two ways: by using the `dredge` method, and the `Anova` method. The results are shown below.

```
BIC_results <- dredge(asset_lm_two, rank = 'BIC')
head(BIC_results, 7)
```

```
## Global model call: lm(formula = fcs ~ factor(survey_year) + asset_qty_poultry +
##     asset_qty_cattle + asset_qty_otherlivestock + asset_qty_sheepgoat +
##     memb_15_44 + hhs_total + bio_bio_12 + house_owned + asset_cartplough +
##     asset_telephone, data = bihs_original, na.action = "na.fail")
## ---
## Model selection table
##      (Int) ass_crt ass_qty_ctt ass_qty_oth ass_qty_plt ass_tlp bio_bio_12
## 291  11.90              4.217                                +
## 292  12.93       +      3.792                                +
## 295  11.90              4.463     -0.4152                    +
## 355  21.89              4.410                                +  -0.003135
## 1315 11.52              4.007                                +
## 419  12.16              3.599                                +
## 299  11.94              4.065                   0.0638       +
##      fct(srv_yer) hhs_ttl mmb_15_44 df    logLik    BIC delta weight
## 291                 6.681            5 -1249.919 2528.4  0.00  0.590
## 292                 6.410            6 -1248.453 2531.1  2.77  0.148
## 295                 6.681            6 -1249.144 2532.5  4.15  0.074
## 355                 6.229            6 -1249.298 2532.8  4.46  0.063
## 1315                6.634    0.6452  6 -1249.639 2533.5  5.14  0.045
## 419             +   6.633            6 -1249.664 2533.6  5.19  0.044
## 299                 6.673            6 -1249.874 2534.0  5.61  0.036
## Models ranked by BIC(x)
```

BIC reports that cattle, telephone and household total are important predictors of food consumption score.

It is important to note that the IC scores above are really close to each other. We did not consider model averaging because we were more interested in looking at which predictors are better at explaining the variation in the response variable `fcs` rather than getting the most accurate estimate of `fcs`. We chose the model that has the fewest predictors in the BIC output as our best model.

Analysis of Variance

```
Anova(asset_lm_two)
```

```
## Anova Table (Type II tests)
##
## Response: fcs
##                          Sum Sq  Df F value    Pr(>F)
## factor(survey_year)         429   1  1.7523 0.1866389
## asset_qty_poultry            28   1  0.1139 0.7360303
## asset_qty_cattle           1264   1  5.1598 0.0238534 *
## asset_qty_otherlivestock    297   1  1.2138 0.2714952
## asset_qty_sheepgoat         156   1  0.6384 0.4249644
## memb_15_44                  354   1  1.4445 0.2304015
## hhs_total                  2736   1 11.1688 0.0009418 ***
## bio_bio_12                 1193   1  4.8691 0.0281306 *
```

```
## house_owned                  96   1  0.3929 0.5312833
## asset_cartplough           1098   1  4.4801 0.0351487 *
## asset_telephone            3157   1 12.8854 0.0003891 ***
## Residuals                 70563 288
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA output above reports that cattle, household total, precipitation, cartplough, and telephone were significant at at least a 0.05 significance level.

Descriptive Statistics

Livestock Assets

```
year_11 <- filter(bihs, survey_year == "2011")
summary(year_11$livestock)%>%
  pander::pander()
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	0	0	0	0	0

```
year_15 <- filter(bihs, survey_year == "2015")
summary(year_15$livestock)%>%
  pander::pander()
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	1	1	0.7782	1	1

Technological Assets Summary

Motorbike

```
summary(year_11$motorbike)%>%
  pander::pander()
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	0	0	0.03462	0	1

```
summary(year_15$motorbike)%>%
  pander::pander()
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0	1	1	0.875	1	1	268

Telephone

```
summary(year_11$telephone)%>%
  pander::pander()
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	0	1	0.7385	1	1

```
summary(year_15$telephone)%>%
  pander::pander()
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
1	1	1	1	1	1	52

Television

```
summary(year_11$television)%>%
  pander::pander()
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	0	0	0.2923	1	1

```
summary(year_15$television)%>%
  pander::pander()
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
1	1	1	1	1	1	179

Cartplough

```
summary(year_11$cartplough)%>%
  pander::pander()
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	0	0	0.1615	0	1

```
summary(year_15$cartplough)%>%
  pander::pander()
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0	0	0	0.4694	1	1	235

Radio

```
summary(year_11$radio)%>%
  pander::pander()
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	0	0	0.04231	0	1

```
summary(year_15$radio)%>%
  pander::pander()
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0	0	0	0.1	0	1	274

Tractor

```
summary(year_11$tractor)%>%
  pander::pander()
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	0	0	0.003846	0	1

```
summary(year_15$tractor)%>%
  pander::pander()
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0	0	0	0	0	0	283

Conclusion

Our research question cannot be answered from our study of BIHS dataset above. Although both response variables in both datasets are similar measures of food security, we had so many missing predictors in BIHS dataset that were not present that were used in the regression for GAC Livelihoods survey. Due to this challenge, we do not think the Differnece-in-Difference analysis can be carried out. Therefore, we cannot conclude whether the GAC Livelihoods programs had a positive impact we can analyze using the BIHS dataset.