# BIHS Model Fitting

STAT-245, Calvin University

*Ivanna Rodriguez, Myungha Kim, James Eapen*

*November 24, 2019*

## Contents



The following document outlines our process for fitting a model that would predict food security score with the BIHS dataset.

We used the variable `fcs` as our response variable. This is the food consumption score in the BIHS dataset that is a similar measure to the food security scores used in the GAC Livelihoods regressions. For our predictors, we looked at the predictors from the GAC Livelihoods that were statistically significant, and tried to find similar predictors in the BIHS dataset, which was a difficult task. We determined that `house_owned` and all of the assets predictors would be good candidates.

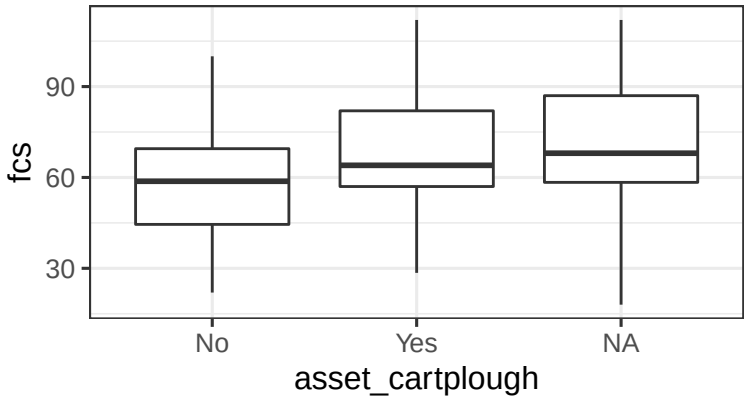
We thought that the best way to go about determining what predictors were significant was to fit a model with as many variables from BIHS as we could, in addition to those found significant in the GAC Likelihoods regressions.

## Exploratory plots

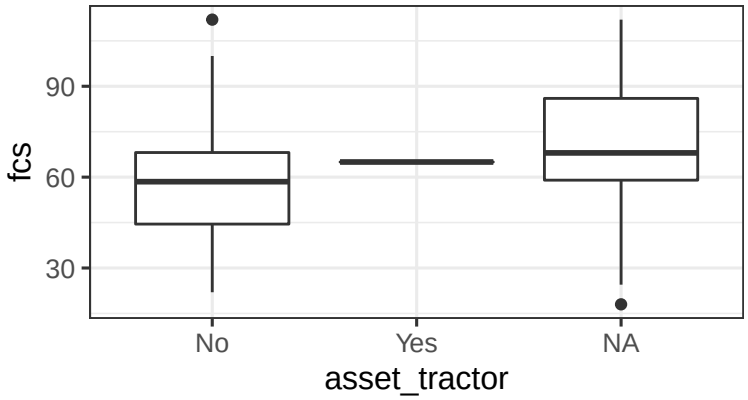
We selected a few categorical variables for the BIHS dataset to see if there might be a relationship between the response `fcs` and the predictors. The boxplots are shown below.

```
# whether they owned/rented a house
gf_boxplot(data = bihs, fcs ~ house_owned)
```

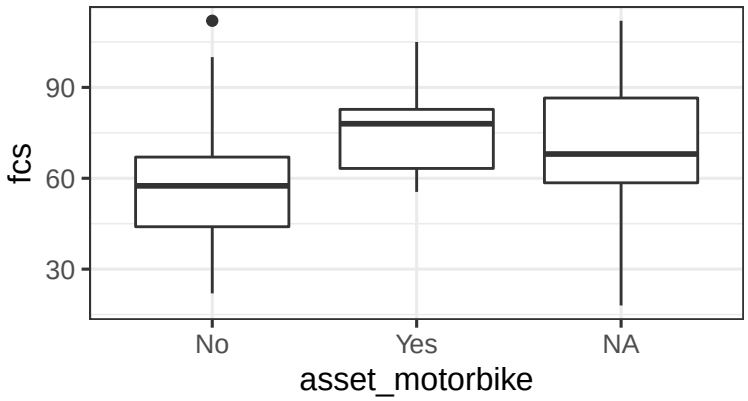

```
# tech assets
gf_boxplot(data = bihs, fcs ~ asset_cartplough)
```



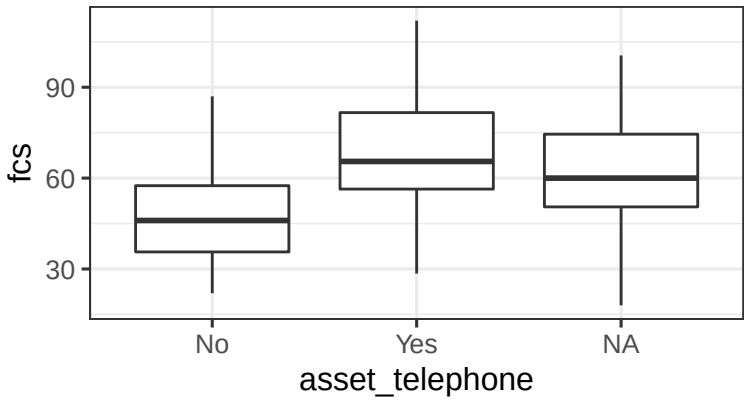

```
gf_boxplot(data = bihs, fcs ~ asset_tractor)
```



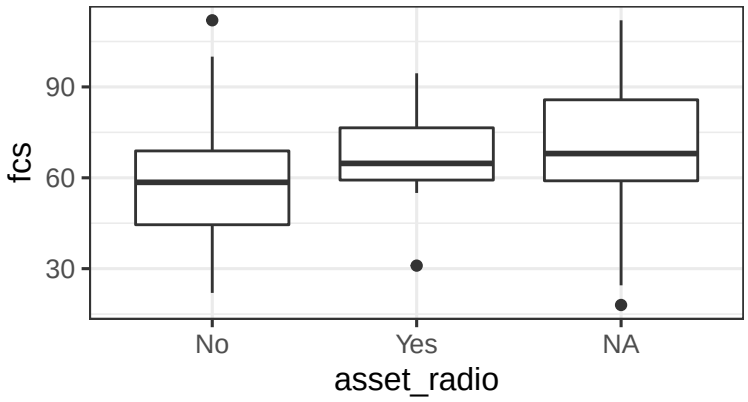

```
gf_boxplot(data = bihs, fcs ~ asset_motorbike)
```




```
gf_boxplot(data = bihs, fcs ~ asset_telephone)
```




```
gf_boxplot(data = bihs, fcs ~ asset_radio)
```




## Fitting the model

We encountered some problems while trying to fit all of our variables into a single model. The model below is one that worked the best for now. It includes our 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')
```

```r
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
```

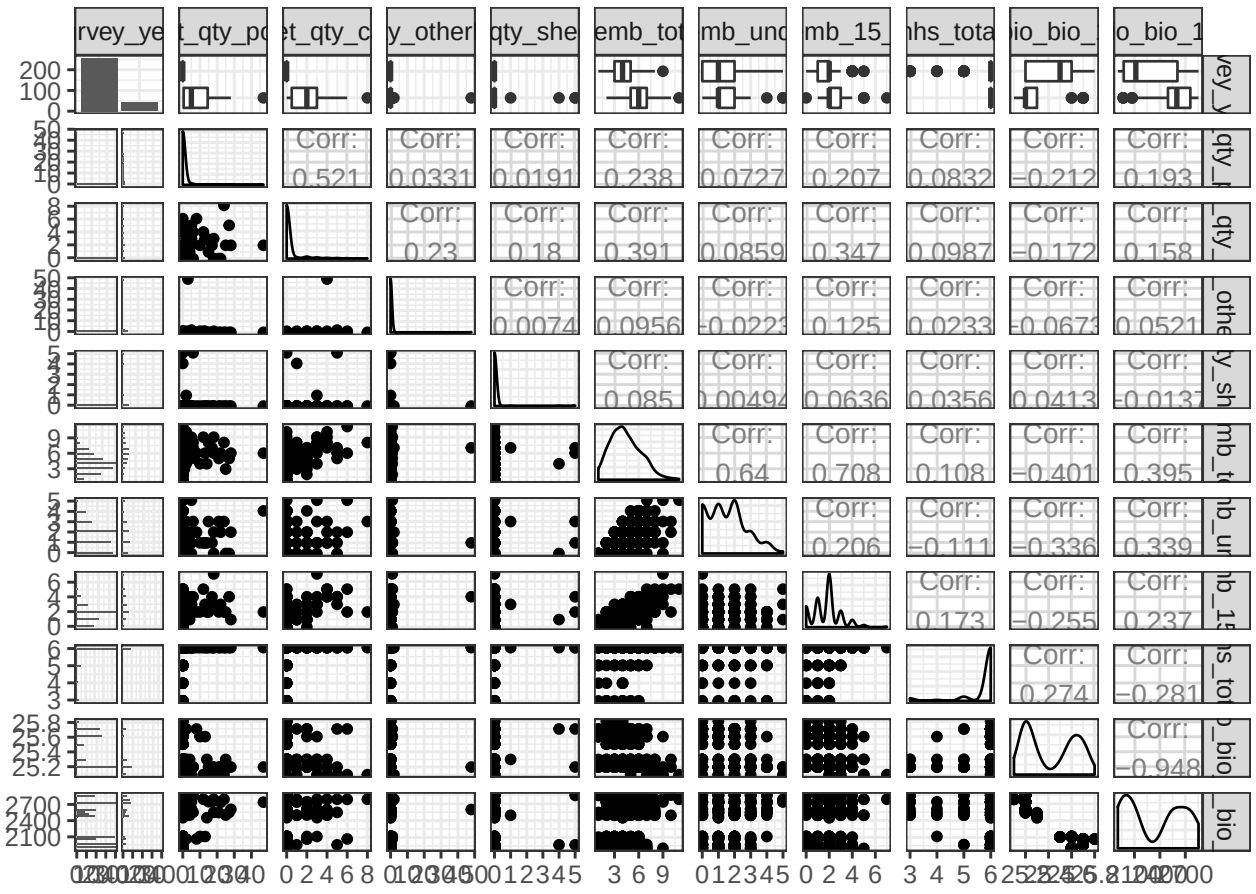
## Checking for multicollinearity between predictors

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

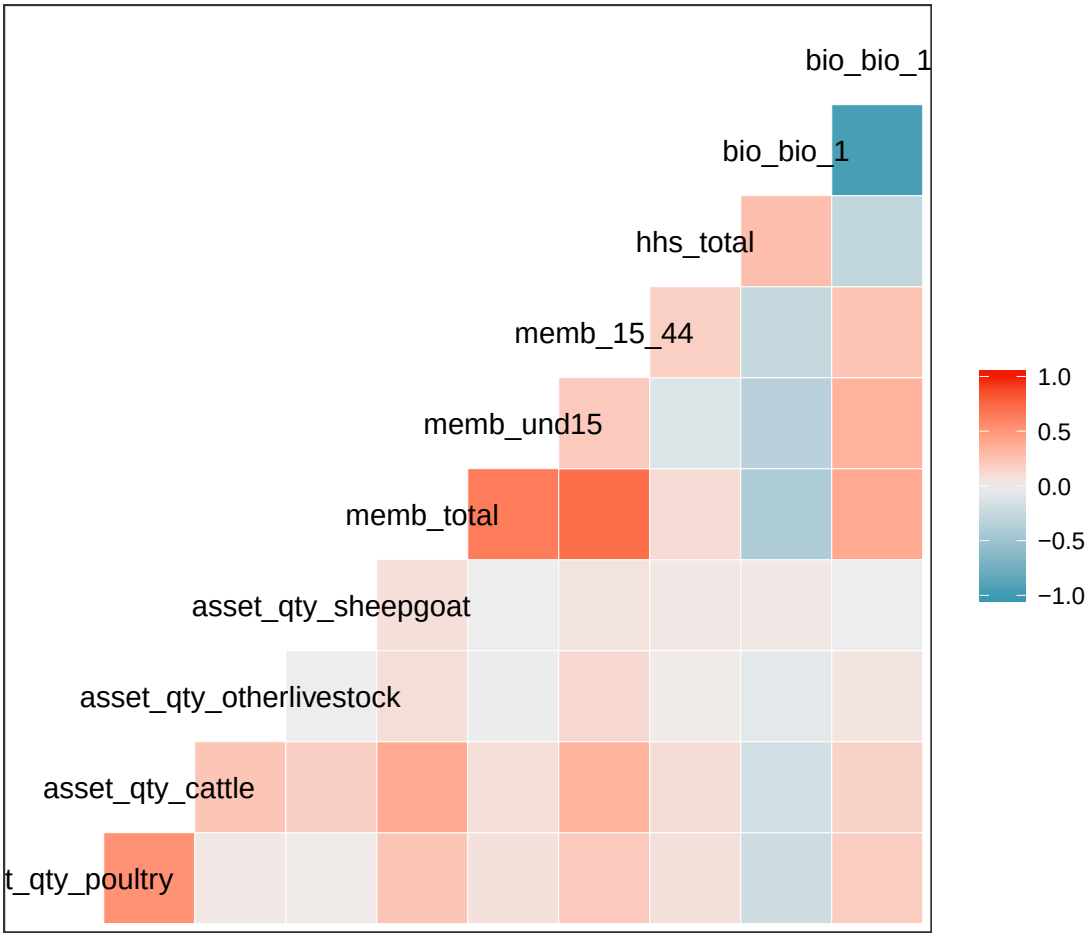
```
##      factor(survey_year)        asset_qty_poultry         asset_qty_cattle
##                 3.175399                 1.785341                 2.411222
## asset_qty_otherlivestock      asset_qty_sheepgoat              memb_total
##                 1.094819                 1.175269                 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)
```




## Removing predictors with high correlation

Because we found correlations between some of the predictors, we decided to remove them and create a new model. The following model is the one we decided to use for this analysis.

```
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
##
```

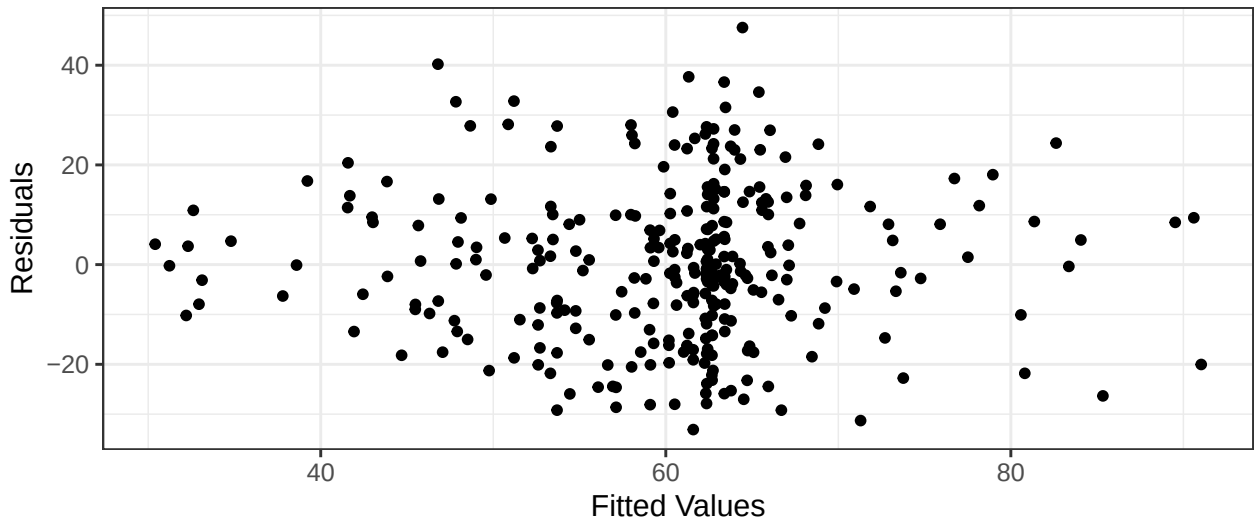
```
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              37.925278  14.041200   2.701 0.007323 **
## factor(survey_year)2015   5.804535   4.384948   1.324 0.186639
## asset_qty_poultry        -0.080570   0.238767  -0.337 0.736030
## asset_qty_cattle          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)
```

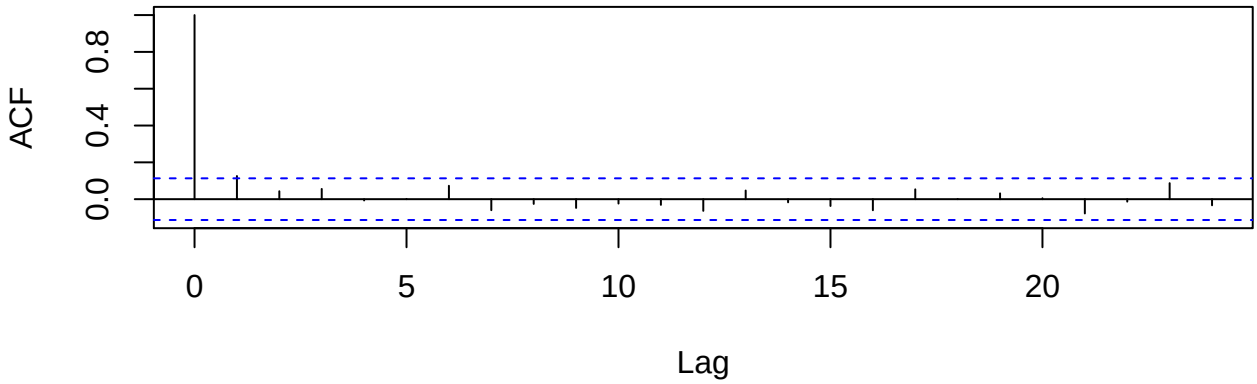
```
##      factor(survey_year)        asset_qty_poultry         asset_qty_cattle
##                 2.890833                 1.732250                 2.392948
## asset_qty_otherlivestock      asset_qty_sheepgoat               memb_15_44
##                 1.088992                 1.163380          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

## Model assessment

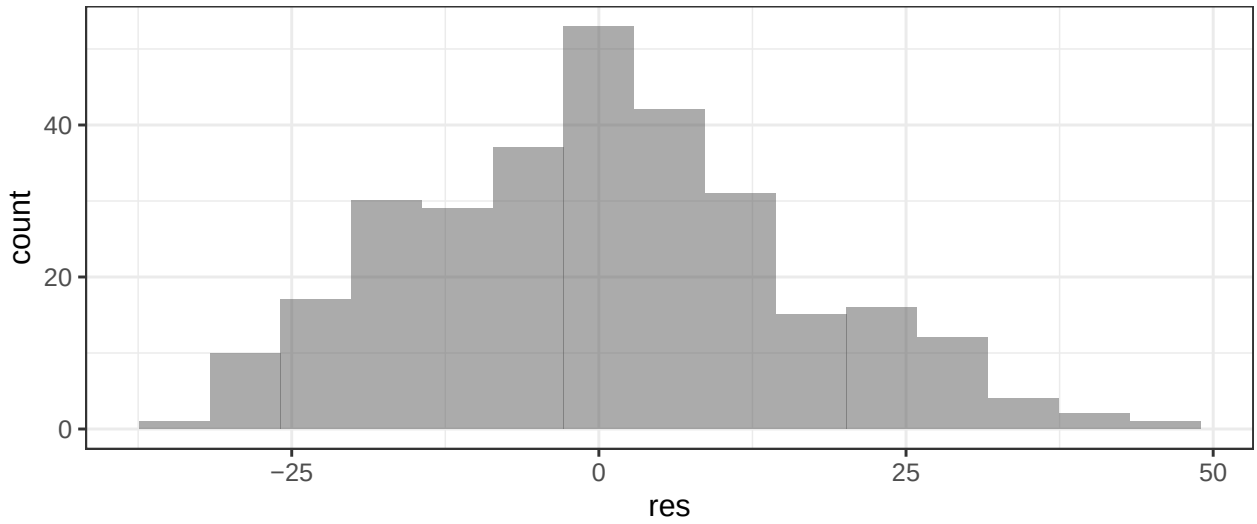
```
# linearity and constant variance
bihs_original <- bihs_original %>%
  mutate(res = resid(asset_lm_two),
         fitted = predict(asset_lm_two))
gf_point(res ~ fitted, data = bihs_original) %>%
  gf_labs(x = 'Fitted Values', y = 'Residuals')
```




```
# 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.

```
AIC_results <- dredge(asset_lm_two, rank = 'AIC')
head(AIC_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_tlp bio_bio_12 fct(srv_yer)
## 356  28.71       +       3.972                   +  -0.004861
## 292  12.93       +       3.792                   +
## 484  31.59       +       3.075                   +  -0.005641            +
## 1380 30.81       +       3.729                   +  -0.005700
## 360  28.18       +       4.190     -0.3339       +  -0.004720
## 1508 34.02       +       2.756                   +  -0.006589            +
## 296  12.86       +       4.033     -0.3601       +
##      hhs_ttl mmb_15_44 df    logLik    AIC delta weight
## 356    5.633           7 -1247.058 2508.1  0.00  0.224
## 292    6.410           6 -1248.453 2508.9  0.79  0.151
## 484    5.446           8 -1246.492 2509.0  0.87  0.145
## 1380   5.452    0.9344 8 -1246.508 2509.0  0.90  0.143
## 360    5.673           8 -1246.554 2509.1  0.99  0.137
## 1508   5.240    0.9983 9 -1245.864 2509.7  1.61  0.100
## 296    6.428           7 -1247.871 2509.7  1.63  0.099
## Models ranked by AIC(x)
```

```
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)
```

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

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

It is imporant to note that the IC scores above are really close to each other, so we chose the models that had the least number of predictors to make a decision on which was the best model, but this is subjective.

## 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 reports that cattle, household total, precipitation, cartplough, and telephone were significant at different significance levels which are outlined in the output above.