Stats 101A
Project Report

Langkun Guo
Eugene Ji
Shiyu Ji
Yuxiao Liu
Enjie Ma
Jiashu Miao
Iris Yang

Introduction

We obtain a dataset from the "General Social Survey 2016" to find the significant predictors that may explain people's happiness levels and present their relationship in the most appropriate regression model. Our method is to fill in the missing data, find the full model, transform the model, select the significant variables, and conclude the best linear model based on our interest.

Methods and Procedures

1. Data cleaning

The dataset has 2361 observations and 13 variables. The last variable is the level of happiness and the first twelve variables are potential predictors of the happiness levels. We start with data cleaning to ensure all values we use to model the relationship between predictors and the response variable are potentially meaningful. We eliminate all observations with happiness levels unknown or not answered because they cannot be used to estimate the predictors' effect. We then identify lots of invalid entries with types of NA, Inapplicable, Don't Know, Not Answered, Can't Choose, and Other. We substitute all these invalid entries with NAs because they may create great leverages and interfere with the reliability of the model.

2. Find the full model

We observe a massive amount of missing values (NAs) in our dataset. Since we want to incorporate as many observations as possible in our model for better predictive power and for a more representative sample group, we need to find a method to fill in these missing values. We intend to replace NAs with valid values according to the distribution of the corresponding predictors, thus we categorize observations into three subgroups according to their happiness levels from 1 to 3 for further use of distribution within each subgroup. Then, we conduct 200 simulations of filling-in NAs and finding a linear regression model for each filled dataset. For each simulation, we fill in NAs with randomly selected valid values from the corresponding columns within each subgroup. For example, observation 1 has a happiness level of 1 and it has an NA value under the variable "Health". We randomly select a valid entry under the column "Health" within the group of happiness level 1 and replace the NA with it.

According to the Maximum Likelihood Estimation, in each subgroup, the distribution of observations in each predictor can be estimated by valid observed values in the corresponding subgroup. We believe this replacement will preserve the original data randomness within each subgroup. We use this method rather than filling in NAs with the means or medians because an increased amount of mean or median values will make the distribution of the predictors' curve to the middle, which may result in overfitting the data. We may go through a data transformation for fixing assumption violations later, thus, to ensure that the model can go through the Box-Cox transformation without much varying the model, we also perform a few variables treatments for variables with a meaningful 0 value: We add 1 to each value under "Children" and change 0 to 0.1 for values under "Working Hours" and "Education".

The next step within each simulation is to find the linear regression model with each dataset after filling in NAs. Different models will be produced by different filled datasets since the values that replace NAs are randomly selected. We need simulations to ensure we obtain a stable result. Our observation indicates that our method is valid and the result is stable because coefficients for each predictor roughly follow normal distributions with small standard deviations **(Figure 1)**.

After checking the validity of our method, we need to choose a full model from all 200 models produced by each simulation. We observe that the adjusted R^2 for different models also follow a normal distribution, concluding a stable outcome. To have the most representative case, we choose the model that

is the closest to the center **(Figure 2)**, which is the mean of the adjusted R-squared value, in our case, 0.2036.

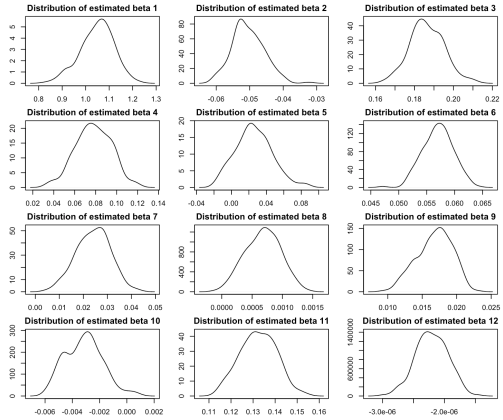

Figure 1: Distribution of coefficients of the models based on our filling method

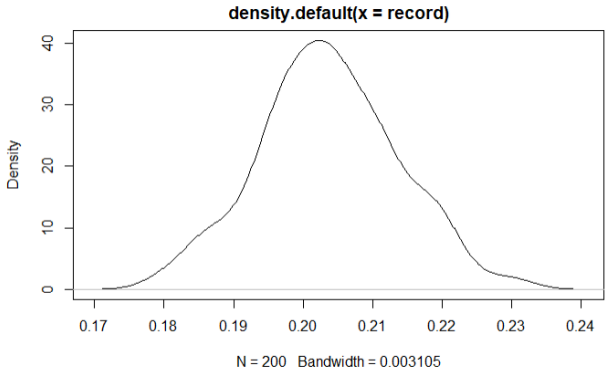

Figure 2: Distribution of Adjusted R-square of the models based on our filling method

3. Data transformation and variable selections

We observe many assumption violations of our full model from the diagnostic plots. The error term is not following a normal distribution and the variation is not constant. There are also many leverages, outliers, and influential points. Therefore, it seems to be necessary to perform data transformation (transformed model in Appendix **Figure A**) to fix the problems in model assumptions. However, there is no much improvement after the transformation as the normality and constant variance are still violated and many leverages still present (shown in **Figure 3-6**).

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.821e-01  1.375e-01   7.141 1.23e-12 ***
## Household   -4.877e-02  1.588e-02  -3.070  0.00217 **
## Health       1.842e-01  1.491e-02  12.353  < 2e-16 ***
## OwnHome      1.058e-01  2.438e-02   4.339 1.49e-05 ***
## Instagram    7.017e-02  2.638e-02   2.659  0.00788 **
## Marital      5.692e-02  8.700e-03   6.542 7.41e-11 ***
## Sex          2.008e-02  2.444e-02   0.822  0.41134
## Age          9.688e-04  8.181e-04   1.184  0.23645
## Children     1.179e-02  8.213e-03   1.436  0.15114
## Education   -4.369e-03  4.355e-03  -1.003  0.31583
## JobSat       1.325e-01  9.884e-03  13.403  < 2e-16 ***
## Income      -2.110e-06  4.258e-07  -4.955 7.74e-07 ***
## WorkHrs     -4.032e-03  9.014e-04  -4.473 8.09e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5829 on 2348 degrees of freedom
## Multiple R-squared:  0.2077, Adjusted R-squared:  0.2036
## F-statistic: 51.29 on 12 and 2348 DF,  p-value: < 2.2e-16
```

Figure 3: Summary of the Full Model

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.0833202  0.2545151  -4.256 2.16e-05 ***
t_Household  0.3064898  0.0838548   3.655 0.000263 ***
t_Health     0.3412387  0.0284235  12.006  < 2e-16 ***
t_OwnHome   -0.1033168  0.0239828  -4.308 1.72e-05 ***
t_Instagram  0.0041208  0.0015827   2.604 0.009283 **
t_Marital    1.1765933  0.2059335   5.713 1.25e-08 ***
Sex          0.0089291  0.0155092   0.576 0.564855
t_Age        0.0133024  0.0213618   0.623 0.533530
t_Children   0.0187968  0.0143703   1.308 0.190992
t_Education -0.0009532  0.0012783  -0.746 0.455948
t_JobSat     0.8154016  0.0643375  12.674  < 2e-16 ***
t_Income    -0.0331119  0.0058613  -5.649 1.81e-08 ***
WorkHrs     -0.0024915  0.0005712  -4.362 1.34e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3689 on 2348 degrees of freedom
Multiple R-squared:  0.2083,    Adjusted R-squared:  0.2042
F-statistic: 51.47 on 12 and 2348 DF,  p-value: < 2.2e-16
```

Figure 4: Summary of the Transformed Model

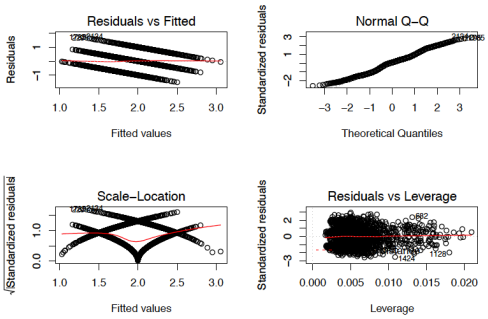

Figure 5: Diagnostic plots of the Full Model

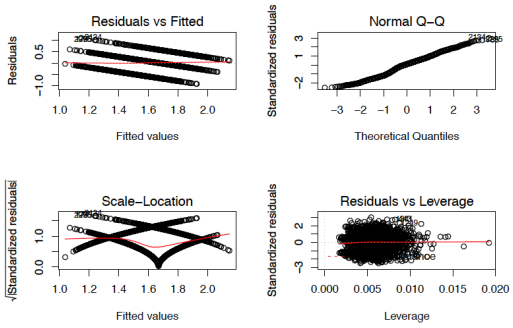

Figure 6: Diagnostic plots of the Transformed Model

We observe many predictors with insignificant p-values in the full model, suggesting that our model has the problem of multicollinearity. Thus we need to perform variable selection to fix the problem. We implement both forward and backward variable selections based on the full model using AIC and BIC **(See Appendix Figure B-E)**. Both stepwise selections using AIC and BIC suggest a model with 9 variables. To compare the reduced model with the full model, we perform the partial F-test **(Figure 7)**, and the resulted p-value suggests that the reduced model is preferred. Therefore we have the reduced model as our final model.

```
## Analysis of Variance Table
##
## Model 1: Happy ~ Household + Health + OwnHome + Instagram + Marital +
##     Children + JobSat + Income + WorkHrs
## Model 2: Happy ~ Household + Health + OwnHome + Instagram + Marital +
##     Sex + Age + Children + Education + JobSat + Income + WorkHrs
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1   2351 798.78
## 2   2348 797.72  3     1.061 1.041 0.3733
```

Figure 7: Partial F-test between the final model and the full model

Side note: Filling in the missing value in the testing dataset:

The model does not recognize missing values if it encounters any of the testing dataset. Therefore we should treat the missing values first before predict the happiness levels with our model. First, all invalid entries, including Inapplicable, Don't Know, Not Answered, Can't choose, and others should be substituted with NA. Then, fill in NAs with the mean of all valid values of the corresponding predictors. Also, add 1 to each value under "Children" and change 0 to 0.1 for values under "Working Hours" and "Education".

We perform cross-checking to ensure the method of filling in NAs in the testing dataset is significant for our constructed model. We randomly select 2000 and 361 observations and name them training set and testing set, respectively. Apply the model produced in the training set to the testing set and perform an F-test. We repeat the procedure for 100 times and record the F-statistics as well as the p-values. The mean F-statistic is 13.43 and the mean p-value is $1.513*10^{-17}$. Thus it's reasonable to believe that this filling method is valid for our model.

Results and interpretation

The R result of our final model (**Figure 8-9**) suggests that all selected variables are significant. The R-squared (0.2024) is slightly increased compared to the full model. The p-value is also significant. However, assumption violations are not improved as suggested by diagnostic plots. But we still observe some linear patterns, normality to some extents, and no leverage points. According to the codebook, a larger happiness value indicates that the person is less happy. Therefore, a positive coefficient suggests that a lower value of the input of the predictor enhances the happiness level. However, the values in each predictor have different meanings according to the codebook. Thus we interpret the meaning of predictors below (Note that Happiness Level 1: Very Happy; Happiness Level 2: Pretty Happy; Happiness Level 3: Not too happy):

- When the other predictors are fixed, one unit increase in household member will decrease the happiness level for 0.05627 unit.
- When the other predictors are fixed, one unit increase in health level (i.e. less healthy) will increase the happiness level for 0.1907 unit.
- When the other predictors are fixed, owning a house, compared to renting a house, will decrease the happiness level for 0.104 unit.
- When the other predictors are fixed, having at least one Instagram account will decrease the happiness level for 0.07426 unit.
- When the other predictors are fixed, being married, compared to who are not, will decrease the happiness level for 0.05433 unit.
- When the other predictors are fixed, having one more child will increase the happiness level for 0.0165 unit.
- When the other predictors are fixed, one unit increase in Job Satisfaction level (i.e. less satisfied with their jobs) will increase the happiness level for 0.1322 unit.
- When the other predictors are fixed, every 10,000 U.S. dollar increase in Salary will decrease the happiness level for 0.0255 unit.
- When the other predictors are fixed, every 10 hours increase in working time will decrease the happiness level for 0.0411 unit.

$$\hat{Happy} = 1.001 - 0.05437 * Household + 0.1877 * Health + 0.1034 * OwnHome + 0.07328 * Instagram + 0.05433 * Marital + 0.0165 * Children + 0.132 * JobSat - 0.000002 * Income - 0.004084 * WorkHrs$$

Figure 8: Regression Model of the Final Model

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.001e+00  9.173e-02  10.917  < 2e-16 ***
Household   -5.437e-02  1.493e-02  -3.641 0.000277 ***
Health       1.877e-01  1.470e-02  12.771  < 2e-16 ***
OwnHome      1.034e-01  2.407e-02   4.297 1.80e-05 ***
Instagram    7.328e-02  2.614e-02   2.803 0.005102 **
Marital      5.433e-02  8.112e-03   6.697 2.65e-11 ***
Children     1.650e-02  7.699e-03   2.143 0.032242 *
JobSat       1.320e-01  9.870e-03  13.372  < 2e-16 ***
Income      -2.206e-06  4.205e-07  -5.245 1.70e-07 ***
WorkHrs     -4.084e-03  9.006e-04  -4.535 6.04e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5829 on 2351 degrees of freedom
Multiple R-squared:  0.2066,    Adjusted R-squared:  0.2036
F-statistic: 68.03 on 9 and 2351 DF,  p-value: < 2.2e-16
```

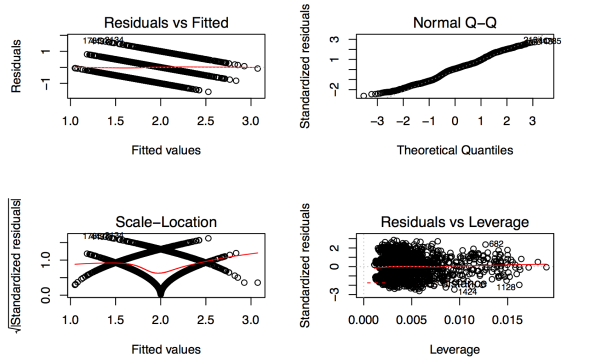

Figure 9: R summary and the diagnostic plots of the Final Model

We've also tried several other models before reaching our final model, as presented below.

Model A: Delete all observations with invalid entries including "Don't Know", "No Answer" etc. Model A shown below:

$$\hat{Happy} = \beta_0 - \beta_1 * Household + \beta_2 * Health + \beta_3 * OwnHome + \beta_4 * Instagram + \beta_5 * Marital + x\beta_6 * Sex + \beta_7 * Age + \beta_8 * Children + \beta_9 * Education + \beta_{10} * JobSat - \beta_{11} * Income - \beta_{12} * WorkHrs$$

Figure 10: Regression Model of the Model A

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.001e+00  9.173e-02  10.917  < 2e-16 ***
Household   -5.437e-02  1.493e-02  -3.641 0.000277 ***
Health       1.877e-01  1.470e-02  12.771  < 2e-16 ***
OwnHome      1.034e-01  2.407e-02   4.297 1.80e-05 ***
Instagram    7.328e-02  2.614e-02   2.803 0.005102 **
Marital      5.433e-02  8.112e-03   6.697 2.65e-11 ***
Children     1.650e-02  7.699e-03   2.143 0.032242 *
JobSat       1.320e-01  9.870e-03  13.372  < 2e-16 ***
Income      -2.206e-06  4.205e-07  -5.245 1.70e-07 ***
WorkHrs     -4.084e-03  9.006e-04  -4.535 6.04e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5829 on 2351 degrees of freedom
Multiple R-squared:  0.2066,    Adjusted R-squared:  0.2036
F-statistic: 68.03 on 9 and 2351 DF,  p-value: < 2.2e-16
```

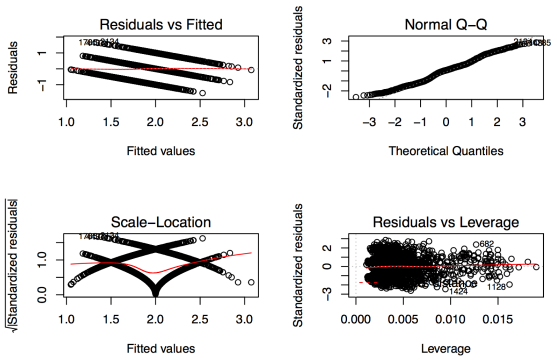

Figure 11: R summary and the diagnostic plots of the Model A

Model B (Full Model): Fill in NAs with random valid values from the same variable groups with variable treatment but without any data transformation or variable selection. Model B shown below (R output and diagnostic plots shown in **Figure 3 and Figure 5** in pp. 3-4):

$$\hat{Happy} = 0.9821 - 0.04877 * Household + 0.1842 * Health + 0.1058 * OwnHome + 0.07017 * Instagram + 0.05692 * Marital + 0.02008 * Sex + 0.00097 * Age + 0.1179 * Children - 0.00437 * Education + 0.1325 * JobSat - 0.000002 * Income - 0.00403 * WorkHrs$$

Figure 12: Regression Model of the Model B

Model C (Final Model): Fill in NAs with random valid values from the same variable groups with variable treatment and variable selection. Model C shown below (R output and diagnostic plots shown in **Figure 8** in pp.6):

$$\hat{Happy} = 1.001 - 0.05437 * Household + 0.1877 * Health + 0.1034 * OwnHome + 0.07328 * Instagram + 0.05433 * Marital + 0.0165 * Children + 0.132 * JobSat - 0.000002 * Income - 0.004084 * WorkHrs$$

Figure 13: Regression Model of the Model C

We compare Model A, B, C before choosing a final model. Model A has a much lower Adjusted R-squared, three insignificant predictors, and assumption violations including non-linear relation, in-normality, non-constant variance, and around 109 leverage points. Model B is our full model before any transformation and Model C is our final model. They are similar and we use the full model as our base model for variable selection. The partial F-test (**Figure 7**) suggest that Model B should be rejected and Model C should be taken. Thus we conclude our final model as Model C.

Discussion

In the interest of finding explanatory variables relating to the happiness levels, we perform a series of data analysis techniques to produce a model. Our method includes filling in missing values, transforming data, selecting variables, and comparing model candidates. After several attempts, we conclude our final model which includes variables household, health, property ownership (OwnHome), Instagram, marital status, job satisfaction, income, and working hours.

Many researchers have reached similar conclusions on the significance of variables in our final model. At the household level, Sehee Han (2015) observes that respondents who rent or live in free housing had lower happiness levels compared to those living in their own homes. Also, "household income was associated with happiness". Esteban Ortiz-Ospina and Max Roser's research (2013) support such finding and observe that richer people tend to be happier than poorer people under same social environment. Yukiko Uchida et al. (2013) suggests that the number of household member is also a positive indicator. However, KJ Dell'Antonia suggests that in the United States, non-parents are happier than parents and there exists a parental happiness gap (2016). At the individual level, Han concludes that respondents with a married status are higher in happiness than those with other status. Also, "self-rated health status [is] positively associated with happiness". Furthermore, Derek Phillips (1967) discovers that happiness levels are conditional on people's state of mental health after interviewing 600 adults from the state of New Hampshire. Researchers at University of Southern California Marshall School of Business conclude that people taking photos are more likely to feel engaged, more submerged into the community, and happier with life (2016). At the career level, Ronald F. Piccolo et.al.'s (2005) numerical summary suggests a strong, positive intercorrelation between job-satisfaction and happiness. One of Douglas B. Holt's (1998) key findings is that Americans generally work longer hours, which leads to higher a happiness level because of higher income.

Our method is more likely a classification method, which is not a typical way to analyze and conclude a linear regression model. The missing data filling method may not work if the sample size of the non-NA values is small. However, there are many limitations and further improvements. A major limitation of our final model is its assumption violations: high level of normality is not obtained, and the variance is not constant according to the diagnostic plots of the final model.

In addition, there exist some categorical variables, such as Marital, in which there are no internal linear relationships. Besides, the coefficients of the predictors do not directly reflect their relationships with happiness levels since the happiness levels are in a descending order (Happiness Level 1: Very Happy; Happiness Level 2: Pretty Happy; Happiness Level 3: Not too happy). For example, higher number of children means lower happiness levels. Given this limitation, one potential improvement is to change the happiness levels to an ascending order.

Appendix

References

Dell'Antonia, KJ. "For U.S. Parents, a Troubling Happiness Gap." The New York Times, June 17, 2016. https://well.blogs.nytimes.com/2016/06/17/for-u-s-parents-a-troubling-happiness-gap/. Accessed Mar. 22, 2016

Han, S. "Social Capital and Subjective Happiness: Which Contexts Matter?". *Journal Happiness Study,* vol. 16, issue 1, 2015, pp. 241-255. Springer, doi: 10.1007/s10902-014-9506-7.

Holt, Douglas B. "Does Cultural Capital Structure American Consumption?" *Journal of Consumer Research*, vol. 25, issue 1, 1998, pp. 1-25. doi: 10.1086/209523.

Mikel, Betsy. "The Science of Instagram: How It Actually Makes You Happier." Inc. com, June 13, 2016. https://www.inc.com/betsy-mikel/the-science-of-instagram-how-it-actually-makes-you-happier.htmlAccessed Mar. 22, 2016.

Ronald F. P., Timothy A. J., Koji T., Naotaka W., Edwin A. L. "Core self-evaluations in Japan: relative effects on job satisfaction, life satisfaction, and happiness". *Journal of Organizational Behavior,* 26, 2005, pp. 965-984. doi: 10.1002/job.358

Derek L. P. "Mental Health Status, Social Participation, and Happiness". *Journal of Health and Social Behavior*, vol. 8, no. 4, 1967, pp. 285-291. JSTOR, http://www.jstor.org/stable/2948422

Ortiz-Ospina, E., & Max R. "Happiness and Life Satisfaction." *Our World in Data*, 2013, ourworldindata.org/happiness-and-life-satisfaction.

Uchida, Y., Norasakkunkit, V. & Kitayama, S. "Cultural constructions of happiness: Theory and empirical evidence." In A. Delle Fave (Ed.), *The exploration of happiness*, 2013, pp. 269–280. Netherlands: Springer.

Model and Figures

$$\hat{Happy} = 0.9821 - 0.04877 * Household^{0.7389} + 0.1842 * Health^{-0.2865} + 0.1058 * OwnHome^{-1.6347} + 0.07017 * Instagram^{3.5317} + 0.05692 * Marital^{0.3273} + 0.02008 * Sex + 0.00097 * Age^{0.3273} + 0.1179 * log(Children) - 0.00437 * Education^{1.221} + 0.1325 * JobSat^{0.2074} - 0.000002 * Income^{0.1921} - 0.00403 * WorkHrs$$

Figure A: Regression Model for the Transformed Model

```
Step:  AIC=-2538.76
Happy ~ JobSat + Health + Marital + Income + WorkHrs + OwnHome +
    Household + Instagram + Children

            Df Sum of Sq    RSS     AIC
<none>                    798.78 -2538.8
+ Age        1   0.51004 798.27 -2538.3
+ Education  1   0.34663 798.43 -2537.8
+ Sex        1   0.21666 798.56 -2537.4
```

Figure B: R output: Forward selection with AIC

```
Step:  AIC=-2538.76
Happy ~ Household + Health + OwnHome + Instagram + Marital +
    Children + JobSat + Income + WorkHrs

            Df Sum of Sq    RSS     AIC
<none>                    798.78 -2538.8
- Children   1     1.560 800.34 -2536.2
- Instagram  1     2.670 801.45 -2532.9
- Household  1     4.505 803.28 -2527.5
- OwnHome    1     6.275 805.05 -2522.3
- WorkHrs    1     6.989 805.76 -2520.2
- Income     1     9.347 808.12 -2513.3
- Marital    1    15.239 814.02 -2496.1
- Health     1    55.414 854.19 -2382.4
- JobSat     1    60.748 859.52 -2367.7
```

Figure C: R output: Backward selection with AIC

```
Step:  AIC=-2533.91
Happy ~ JobSat + Health + Marital + Income + WorkHrs + OwnHome +
    Household + Instagram + Children

            Df Sum of Sq    RSS     AIC
<none>                    798.78 -2533.9
+ Age        1   0.51004 798.27 -2532.9
+ Education  1   0.34663 798.43 -2532.4
+ Sex        1   0.21666 798.56 -2532.1
```

Figure D: R output: Forward selection with BIC

```
Step:  AIC=-2533.91
Happy ~ Household + Health + OwnHome + Instagram + Marital +
    Children + JobSat + Income + WorkHrs

            Df Sum of Sq    RSS     AIC
<none>                    798.78 -2533.9
- Children   1     1.560 800.34 -2531.8
- Instagram  1     2.670 801.45 -2528.5
- Household  1     4.505 803.28 -2523.1
- OwnHome    1     6.275 805.05 -2517.9
- WorkHrs    1     6.989 805.76 -2515.8
- Income     1     9.347 808.12 -2508.9
- Marital    1    15.239 814.02 -2491.8
- Health     1    55.414 854.19 -2378.0
- JobSat     1    60.748 859.52 -2363.3
```

Figure E: R output: Backward selection with BIC