Data Haven Well-Being Analysis

Gabriel J. Michael, Ph.D., gabriel.michael@yale.edu

February 25, 2015

We are interested in learning which variables are most important as predictors of well-being, where well-being is defined as a mix of the following:

- Satisfaction with the city or area where you live (yes/no, Question 1)
- Overall health (excellent, very good, good, fair or poor, Question 21)
- Satisfaction with one's work, job, vocation, or daily tasks (completely satisfied, somewhat satisfied, not very satisfied, or not at all satisfied, Question 71)

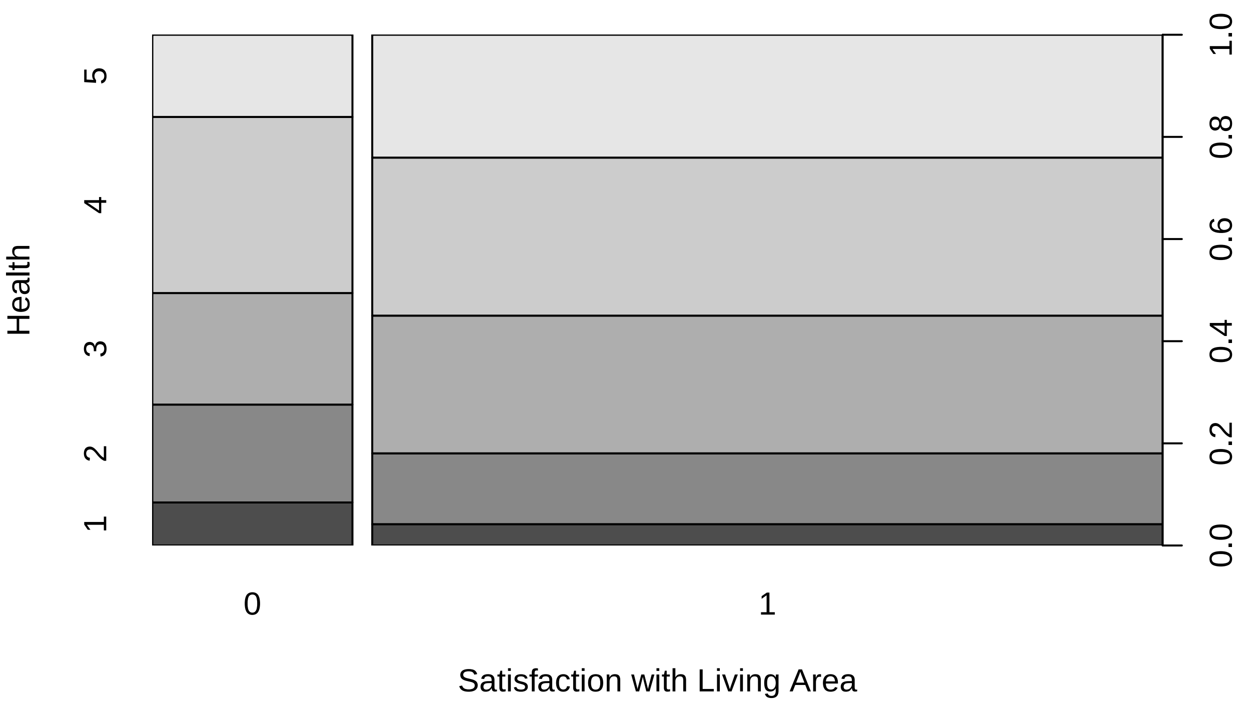
A priori, we would expect each of these questions to have different causal factors, and thus perhaps different useful predictors. Also, predictors may not identify causal factors. Thus, we first want to know the extent to which the responses to each of these three questions correlate. We assess this using measures of correlation for ordinal variables, as well as spine plots, which plot area as a function of the number of responses for a given combination of ordinal variable responses. Note that variables have been recoded so that larger values represent more positive responses.

The correlation between satisfaction with living area and health is small:

```
polychor(dhr$sat_area, dhr$health)
```

```
## [1] 0.1389435
```

```
spineplot(dhr$sat_area,dhr$health, xlab="Satisfaction with Living Area", ylab="Health")
```

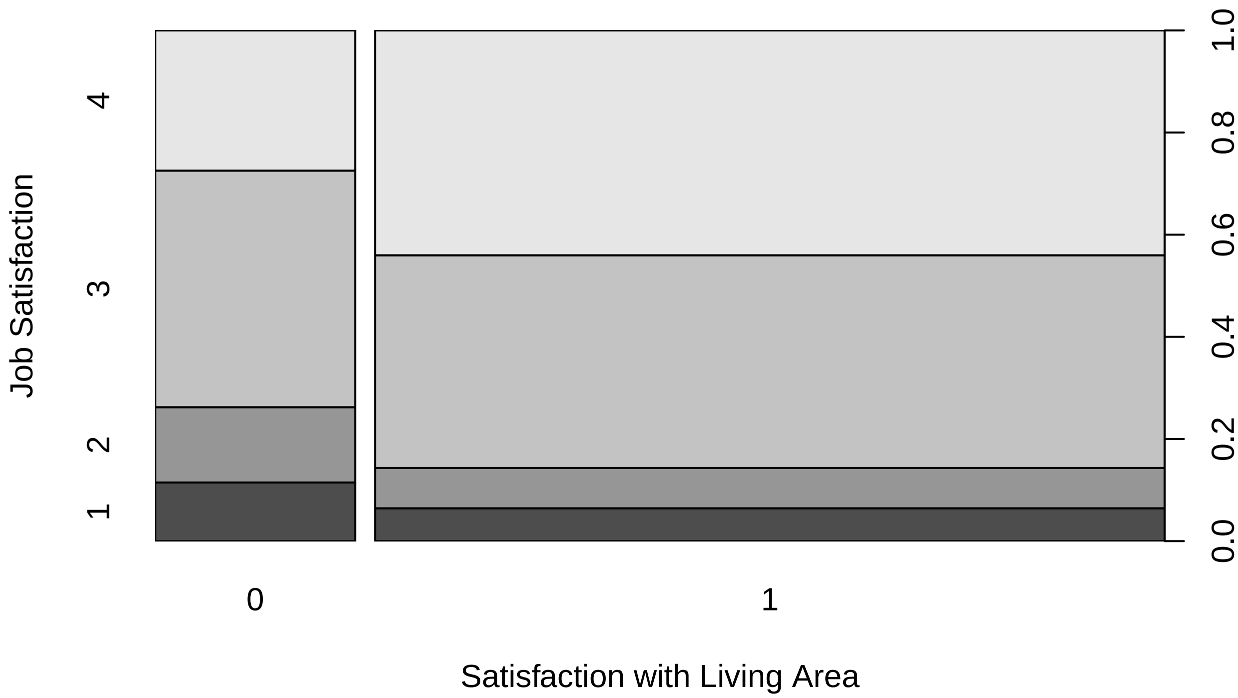


The correlation between satisfaction with living area and job satisfaction is moderate:

```
polychor(dhr$sat_area, dhr$job_sat)
```

```
## [1] 0.2400789
```

```
spineplot(dhr$sat_area,dhr$job_sat, xlab="Satisfaction with Living Area", ylab="Job Satisfaction")
```

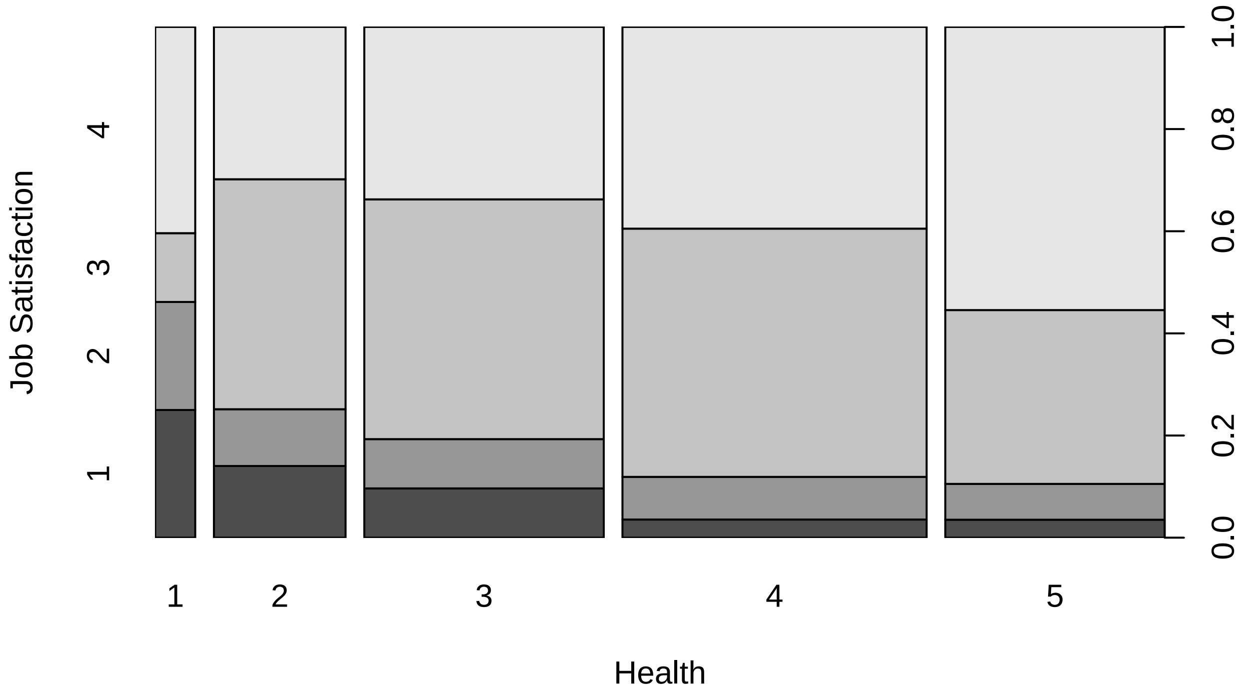


So is the correlation between health and job satisfaction:

```
polychor(dhr$health, dhr$job_sat)
```

```
## [1] 0.253252
```

```
spineplot(dhr$health,dhr$job_sat, xlab="Health", ylab="Job Satisfaction")
```



There is also apparent non-normality in the responses for satisfaction with living area and job satisfaction, with relatively few responses expressing dissatisfaction. This will present a challenge for modeling. In contrast, health responses appear to be more normally distributed. The relatively weak correlations between these three questions suggest that we should not combine them into a single dependent variable.

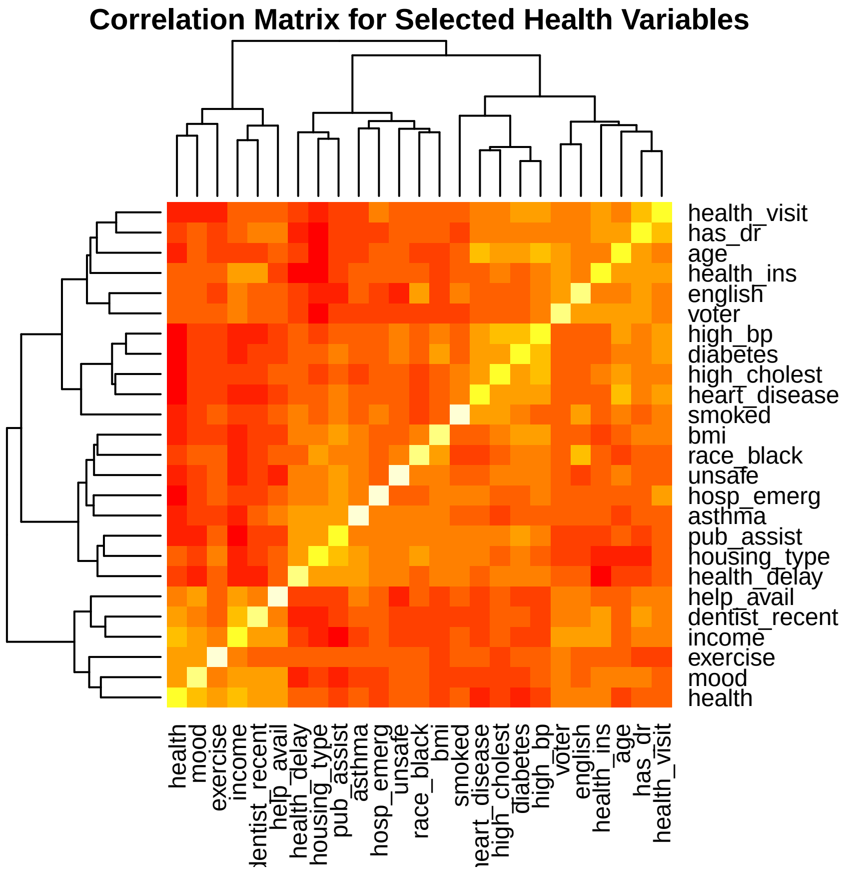
The remainder of this analysis assesses the factors most predictive of responses for each of these three questions.

Health

In this section, we try to predict health as a function of some potentially relevant variables.

There are a large number of variables we might expect to have some value in predicting the health response. This presents a challenging situation - we might expect that many of the potential predictor variables for health are likely to be correlated with one another. If we include all of these variables in a model, we won't be able to distinguish their effects, and effects might be split between highly correlated variables. Let's take a look to see the correlations between the many potentially health-relevant variables.

We can visualize a correlation matrix between many variables using a heatmap, where lighter colors (yellow and white) represent higher correlations:



As it turns out, no two variables have a correlation greater than or equal to 0.70, which is a rough cutoff point for assessing multicollinearity. In fact, the highest correlation is between the indicators for diabetes and high blood pressure, at 0.57.

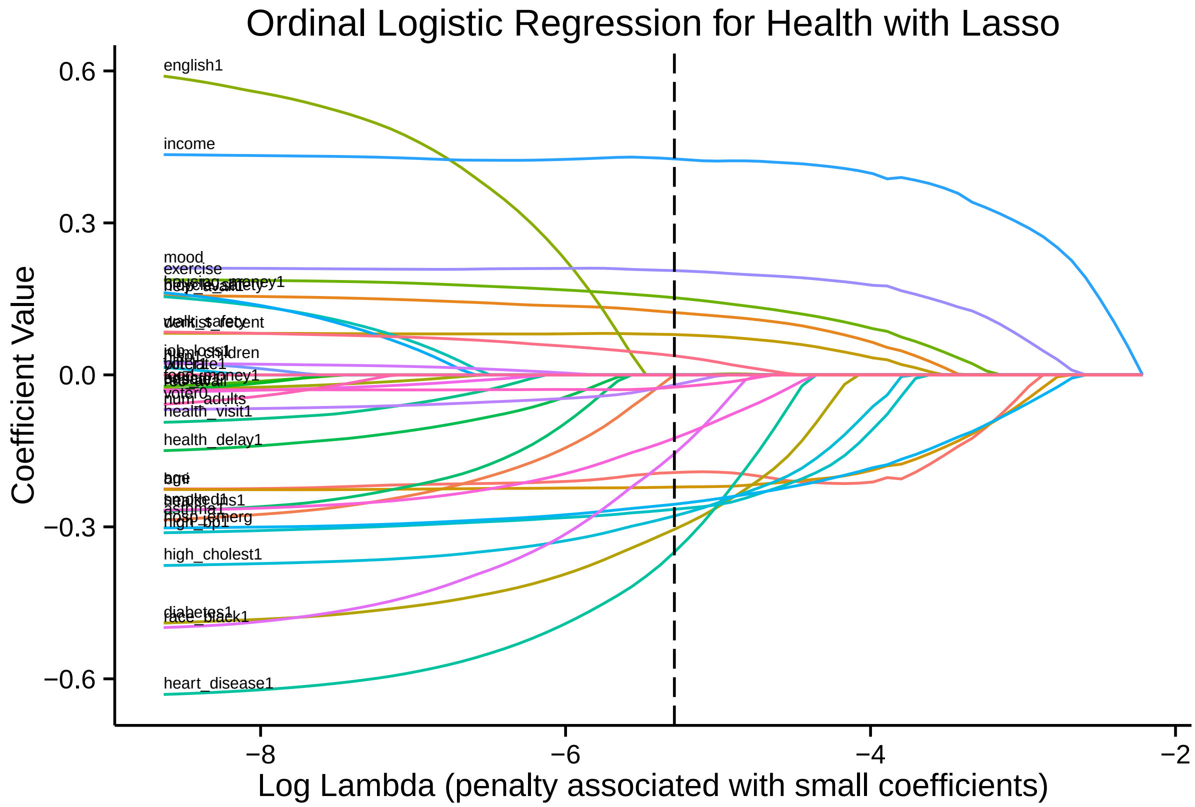
Feature Selection

We still have a large number of variables to choose from. In such a situation, we need a way to select usefully predictive variables while discarding less useful ones. I use the lasso, which eliminates variables when their coefficient estimates are sufficiently small. The following list reports the variables that will potentially be included in the model:

```
##  [1] "income"          "educ"            "voter"           "walk_safety"
##  [5] "bicycle_safety" "rec_avail"       "unsafe"          "age"
##  [9] "bmi"             "high_bp"         "high_cholest"    "diabetes"
## [13] "heart_disease"   "asthma"          "health_ins"      "has_dr"
## [17] "health_visit"    "hosp_emerg"      "dentist_recent" "help_avail"
## [21] "mood"            "exercise"        "food_money"      "smoked"
## [25] "num_children"    "num_adults"      "health_delay"    "job_loss"
## [29] "bill_late"       "housing_money"   "race_black"      "hisp"
## [33] "english"         "health"
```

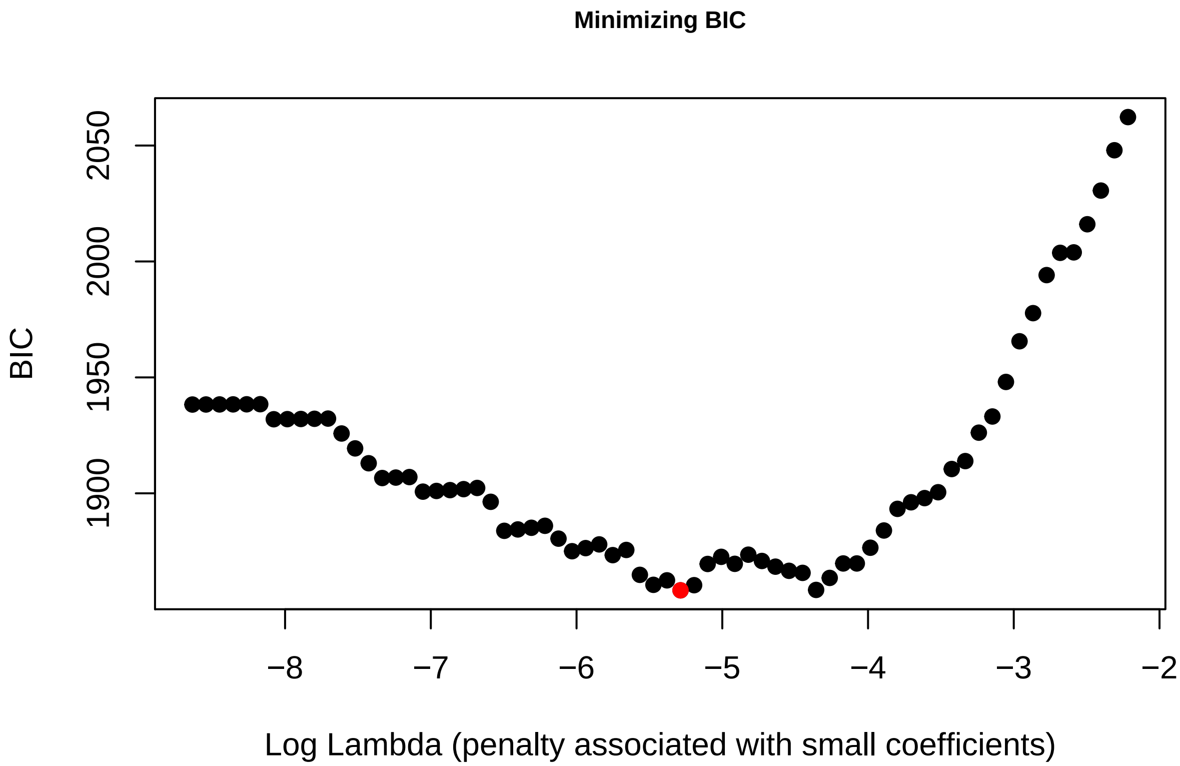
First, I divide the data into a training set and a test set; the training set will be used for model-building, and the test set for model validation. The training set consists of 696 responses randomly selected from the original dataset. Next, I fit a series of continuation ratio models, which are useful for predicting ordinal responses. The following graph shows the coefficient estimates for the series of models with increasing lasso penalties.

```
## Using varname as id variables
```



Larger penalties reduce the number of variables included in the model; thus, at the far right, we see models with only one variable, while at the far left, we see models with all the variables.

The vertical line indicates the penalty I have selected, based on the minimized Bayesian information criterion (BIC). Variables whose lines cross this vertical line will be included in the model; variables whose lines fall short of it will not be included in the model. We want to minimize the BIC, as there are diminishing returns to more complex models. Thus, we select a model with a small BIC and thus a reduced complexity (a larger lambda) as shown in the following plot:



To assess the performance of the chosen model, I first calculate how well it performs on the training data:

```
## [1] 0.4343434
```

On the training data, the model predicts about 43% of the responses correctly. This is significantly better than a random guess (20%), and somewhat better than a naive model that always guesses "Very good" (the most common response, at 31%). Now we test how well it performs on the test data:

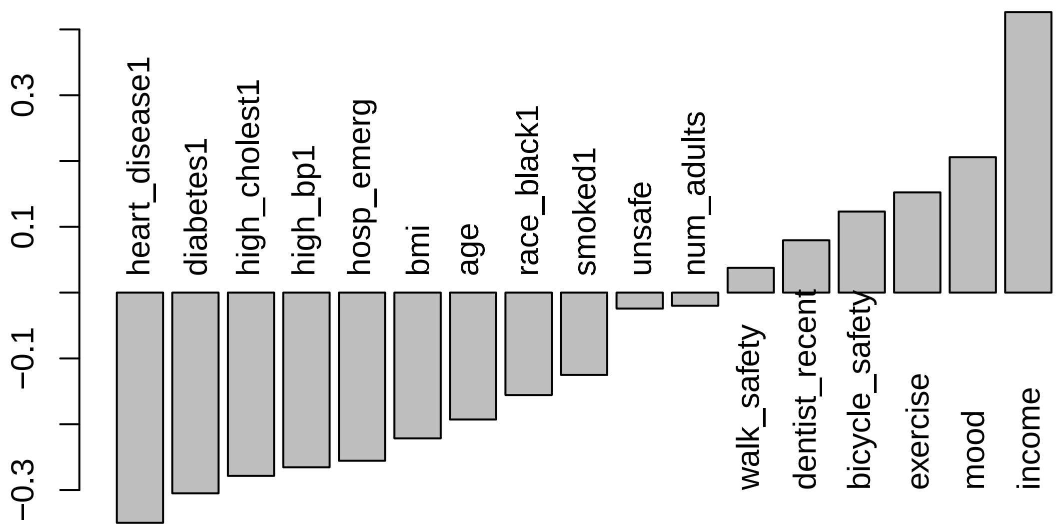
```
## [1] 0.4329004
```

On the test data (232 survey responses randomly selected from the original dataset), the model again predicts about 43% correctly, which suggests that the model has not overfit the training data.

Health Findings

We can now look at the non-zero coefficients for the selected model and get a sense of which variables are most associated with the responses to the health question. The larger the absolute value of a coefficient, the more predictive it is of reported health. Negative values indicate that a response is associated with worse reported health, while positive values indicate the response is associated with better reported health.

Factors most predictive of reported health



Note that with the exception of the dummy variables (yes/no responses), all other variables were standardized prior to running the model. This allows direct comparisons of the coefficients even when variables have drastically different ranges (e.g., income runs from 1 to 6, whereas age runs from 18 to 110). However, this requires assuming that ordinal responses have equidistant intervals, which in some cases is untrue.

Many of the results are not surprising. Increased numbers of visits to hospital emergency rooms, having diabetes, high cholesterol, high blood pressure, asthma, or heart disease, or having been a smoker are all correlated with lower values for reported health. Likewise, older respondents and respondents with larger BMIs also tend to report worse health. Importantly, respondents who indicated their race as black or African American generally reported worse health than those indicating another race.

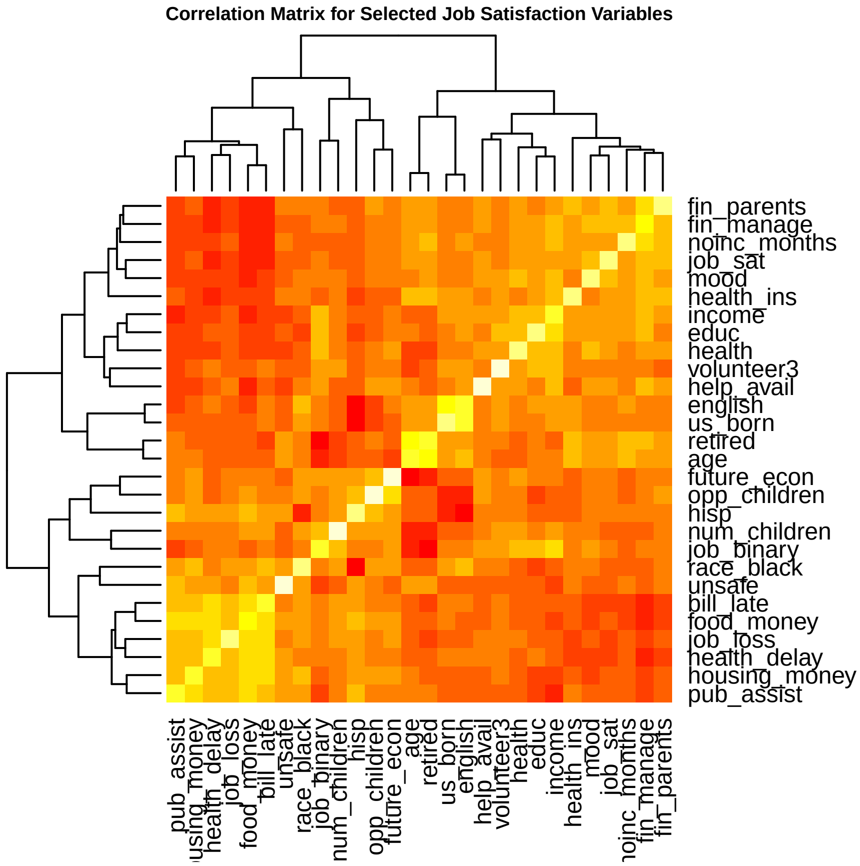
In contrast, respondents with higher incomes report better health, as do those reporting generally positive moods (i.e., only rarely feeling hopeless, down or depressed), those who report more frequent exercise, and those who have recently visited a dentist. Interestingly, respondents who reported having safe places to bicycle in or near their neighborhood also reported better health.

Overall, we have developed a model that performs significantly better than chance or a naive guess, although there is still significant unexplained variance in the responses. Furthermore, it should be emphasized that some predictor variables are likely themselves the results of reported health, rather than causes of it. For example, mood (i.e., how often one feels down or depressed) is highly correlated with reported health, but one's mood is likely influenced by one's health, and vice versa. Likewise, healthy individuals may be more inclined to exercise, which will then help maintain their health.

Job Satisfaction

In this section, we try to predict health as a function of potentially relevant variables. As before, the first step is to identify correlations between predictors.

The correlation matrix helps to identify several highly correlated variables, such as age and being retired, speaking English at home and being born in the U.S., having been late on bills and having had difficulty paying for food, and having been late on bills and having had to put off medical treatment.

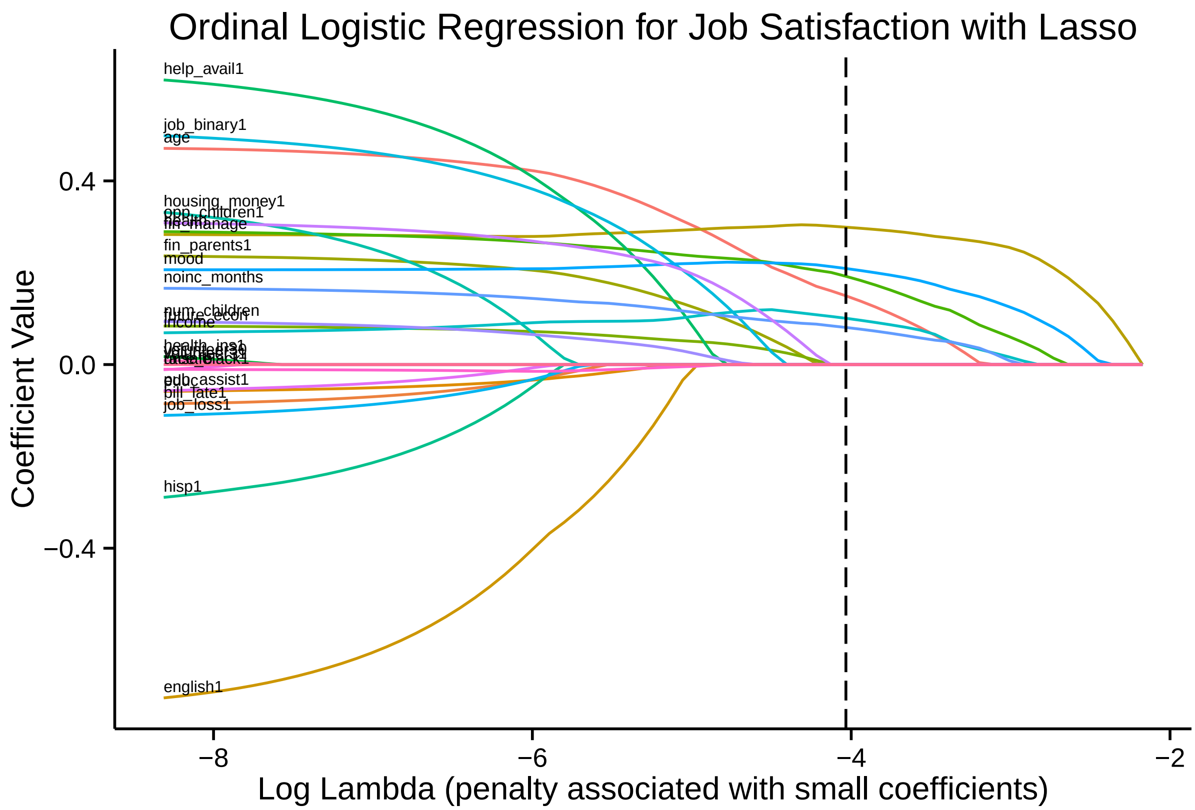


In order to address the issue of multicollinearity, I remove several of these variables (retired, us_born, food_money, and health_delay). After removing these variables, the highest correlation that persists is between education and income, at 0.56. The following list reports the variables that will potentially be included in the model:

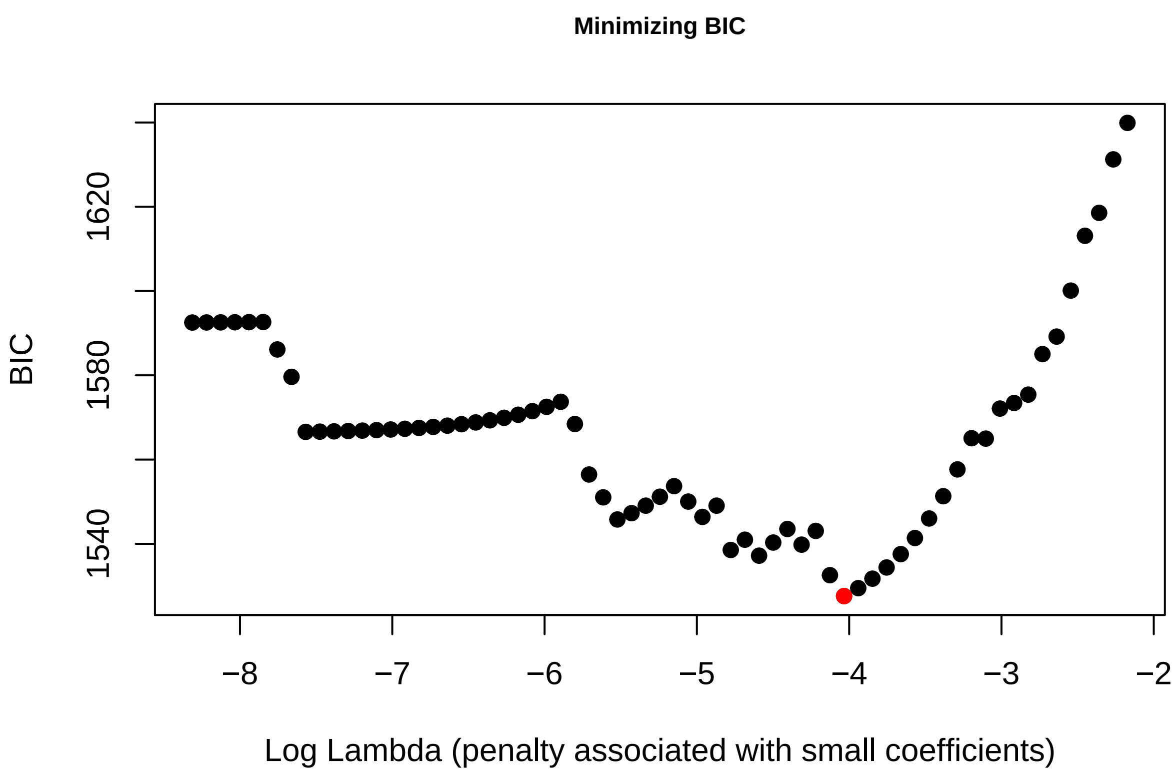
```
##  [1] "volunteer3"    "income"        "unsafe"        "age"
##  [5] "health_ins"    "help_avail"    "mood"          "fin_manage"
##  [9] "future_econ"   "fin_parents"   "opp_children"  "job_loss"
## [13] "bill_late"     "housing_money" "noinc_months"  "num_children"
## [17] "english"       "health"        "race_black"    "hisp"
## [21] "educ"          "pub_assist"    "job_binary"    "job_sat"
```

The training set consists of 702 responses randomly selected from the original dataset. The following graph shows the coefficient estimates for the series of models with increasing lasso penalties.

```
## Using varname as id variables
```



As before, the vertical line indicates the selected penalty, based on the minimized Bayesian information criterion (BIC):



To assess the performance of the chosen model, I first calculate how well it performs on the training data:

```
## [1] 0.508547
```

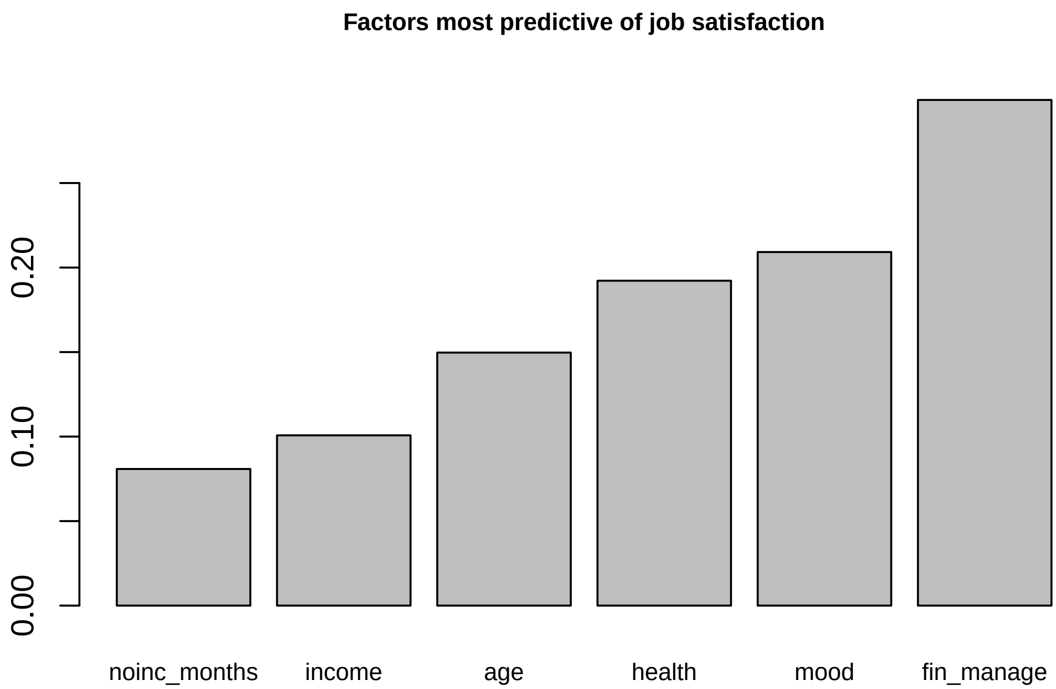
On the training data, the model predicts about 51% of the responses correctly. This is significantly better than a random guess (25%), and somewhat better than a naive model that always guesses "Somewhat satisfied" (the most common response, at 40%). Now we test how well it performs on the test data:

```
## [1] 0.474359
```

On the test data (78 survey responses randomly selected from the original dataset), the model predicts about 47% correctly, which suggests that the model may have slightly overfit the training data by a small amount. Note that this model only improves on a naive model by about 18%.

Job Satisfaction Findings

As before, the absolute value of non-zero coefficients can give us insight into the best predictors of respondents' reported job satisfaction:



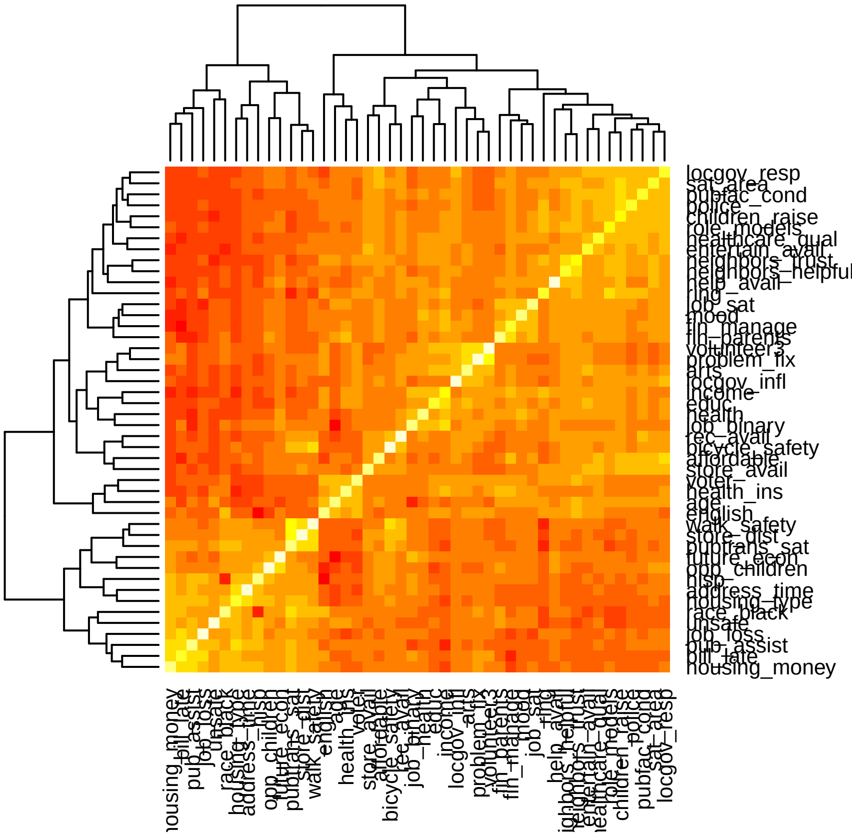
This time, the lasso procedure has selected only a few variables. The strongest effects relate to respondents who reported that were financially managing well, had positive moods, and were in good health. Corresponding to the findings in the Wellbeing Survey Report, both age and income are positively correlated with increases in reported job satisfaction. The number of months a respondent reported being able to live without income is also positively correlated with job satisfaction.

As before, several of these factors are likely endogenous to job satisfaction. That is, one's mood is likely influenced by one's job satisfaction, as is one's ability to manage financially, etc.

Personal Satisfaction

Predicting personal satisfaction, a simple yes/no question (Question 1), is simpler than predicting an ordinal response, but is made more challenging by the preponderance of positive responses. We proceed as before, first identifying potentially problematic correlations:

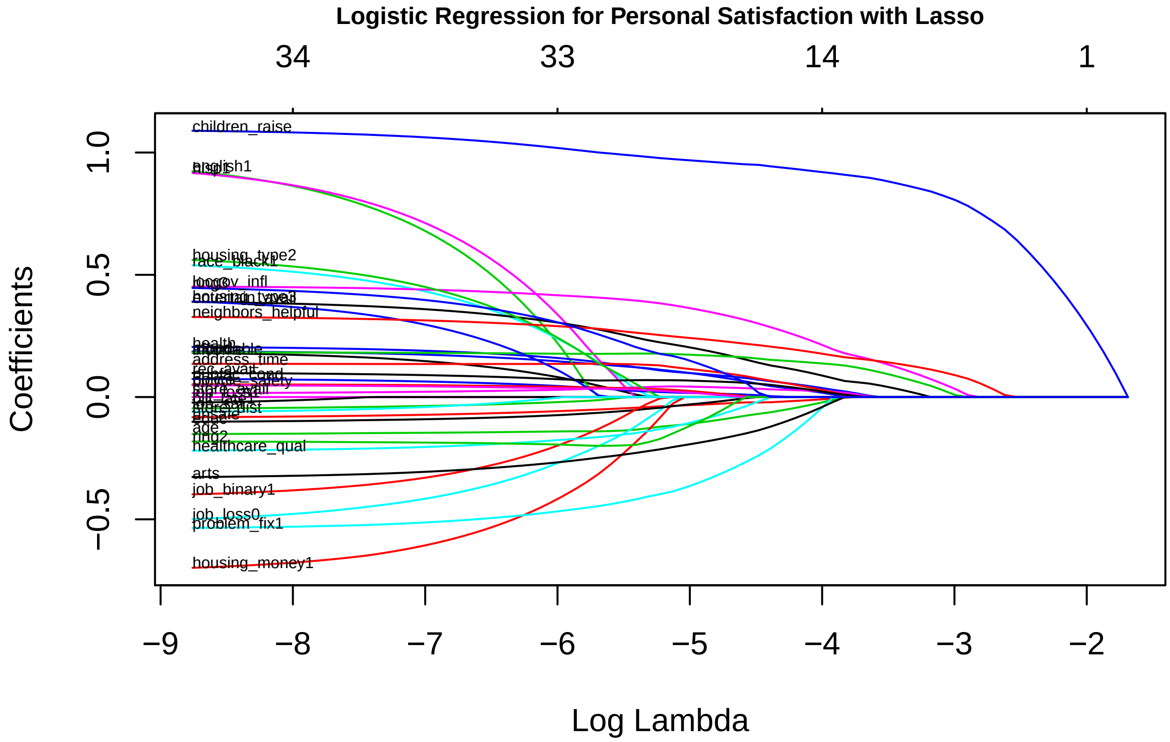
Correlation Matrix for Selected Personal Satisfaction Variables



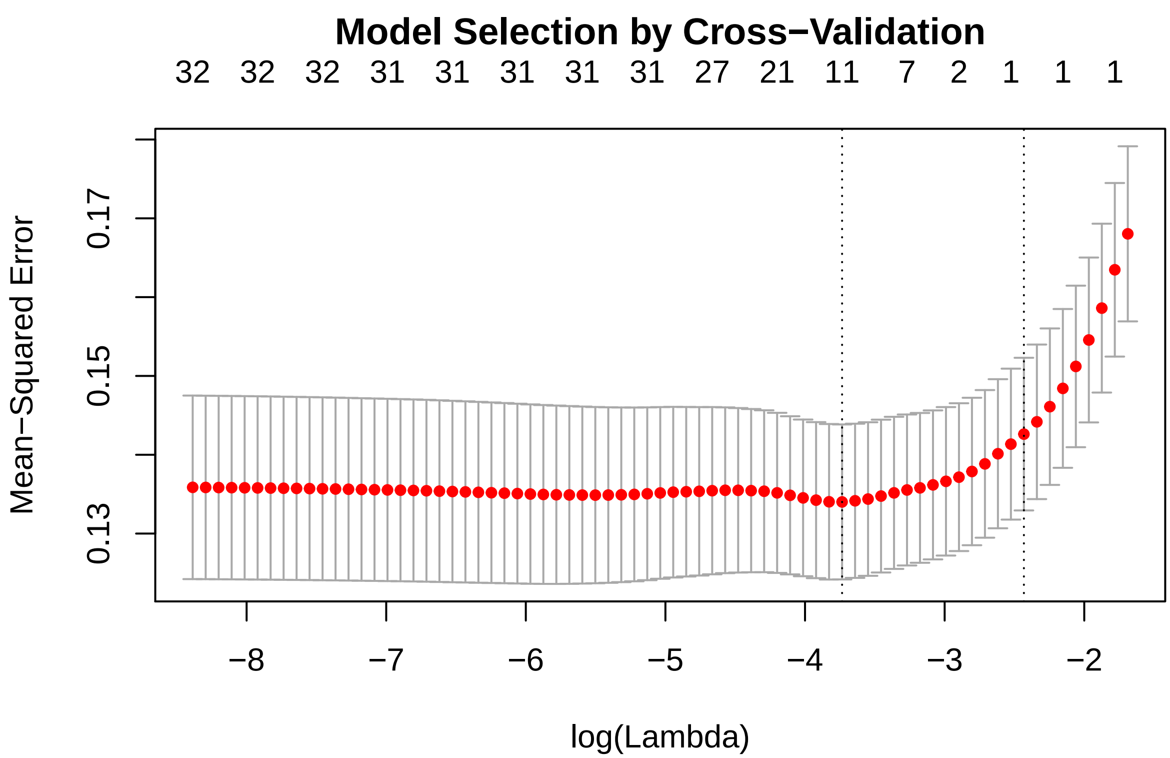
Relatively high correlations exist between the responses for trustworthy and helpful neighbors, so I remove the former variable. A high correlation (0.62) exists between personal satisfaction and whether a respondent views their areas as a good place to raise children, but I retain the latter as a potentially useful predictor of the former. The following list reports the variables that may be potentially included in the model:

```
##  [1] "income"            "unsafe"            "age"
##  [4] "mood"              "job_loss"          "bill_late"
##  [7] "housing_money"     "english"           "health"
## [10] "race_black"        "hisp"              "educ"
## [13] "job_binary"        "job_sat"           "pubfac_cond"
## [16] "healthcare_qual"   "store_avail"       "entertain_avail"
## [19] "police"            "affordable"        "children_raise"
## [22] "problem_fix"       "locgov_infl"       "arts"
## [25] "address_time"      "housing_type"      "store_dist"
## [28] "bicycle_safety"    "rec_avail"         "neighbors_helpful"
## [31] "ring"              "sat_area"
```

The training set consists of 633 responses randomly selected from the original dataset. The following graph shows the coefficient estimates for the series of models with increasing lasso penalties.



With a binomial dependent variable, we are able to use a different procedure to select the appropriate value of lambda. Cross-validation allows us to select a lambda based on multiple random samples of the training data.



To assess the performance of the chosen model, I first calculate how well it performs on the training data:

```
## [1] 0.8025276
```

On the training data, the model predicts about 80% of the responses correctly. This is much better than a random guess (50%). However, it is virtually identical in performance to a naive model that always guesses "Satisfied," since 80% of respondents reported they were satisfied with the area in which they lived.

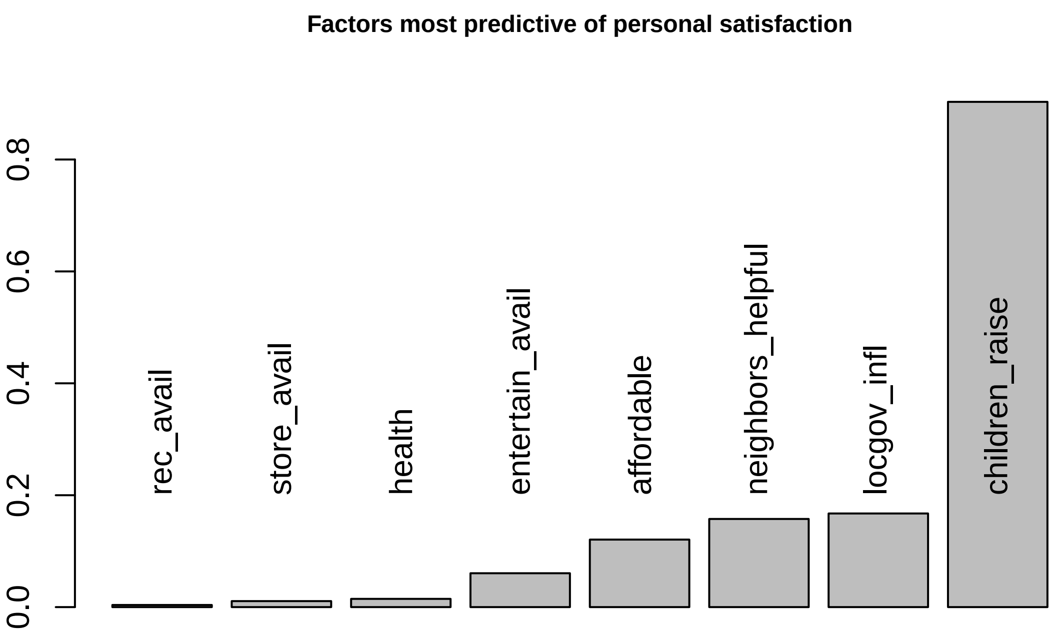
Now we test how well the model performs on the test data:

```
## [1] 0.7857143
```

On the test data (70 survey responses randomly selected from the original dataset), the model predicts about 79% correctly. While this suggests that the model did not overfit the training data, it also is no better than a naive classifier.

Personal Satisfaction Findings

Although the model we developed merely matches the performance of a naive classifier, it can still provide insight into what factors are predictive of personal satisfaction. As before, we examine the non-zero coefficients of the selected model:



The extent to which people believe the Greater New Haven area is a good place to raise children is by far the most predictive factor for personal satisfaction. This should come as no surprise given the relatively strong correlation between the two variables (0.62). Other important factors include the responsiveness of local government, how respondents feel about their neighbors, whether they believe the area is affordable to live in, as well as the availability of various amenities (entertainment, stores, and recreation). Self-reported health also has a measure association with personal satisfaction with one's area.

The Wellbeing Survey Report provides a list of twelve life aspects, ranked in descending order of reported quality. Several of the factors identified above are included in this list. For example, the availability of stores and entertainment are predictive of personal satisfaction, and respondents in aggregate report that these items are generally "good" in the Greater New Haven area. However, affordability and the perceived ability to influence local government are also predictive of personal satisfaction, but respondents in aggregate report that these items are only "fair" in the Greater New Haven area.

Correlated Variables, Homogenous Variables, and Variables with Low Response Rates

Correlated Variables

One of a pair of highly correlated variables might be removed from the survey without losing much information. The following list reports pairs of variables with correlations >= 0.5, along with their numeric correlation.

```
##          us_born           english        food_money         bill_late
##             0.84              0.84              0.76              0.76
##  neighbors_trust neighbors_helpful         bill_late      health_delay
##             0.69              0.69              0.62              0.62
##   children_raise          sat_area       role_models    children_raise
##             0.62              0.62              0.58              0.58
##         diabetes           high_bp        food_money      health_delay
##             0.57              0.57              0.56              0.56
##  neighbors_trust    children_raise      high_cholest           high_bp
##             0.54              0.54              0.53              0.53
##  neighbors_trust       role_models      health_visit            has_dr
##             0.53              0.53              0.52              0.52
##      role_models          sat_area
##             0.50              0.50
```

Homogenous Variables

Variables with largely homogenous results (i.e., low variance) might not provide particularly useful information; the following list reports variables with a variance <= 0.10, along with their numeric variance. In particular, school_act, housing_money, english, and help_avail exhibit very low variance. One might want to retain english for descriptive statistical purposes, but the others could likely be discarded without losing much information.

```
##    school_act housing_money           q6r       english    help_avail
##          0.03          0.06          0.06          0.06          0.07
##    health_ins        has_dr          hisp  health_visit
##          0.08          0.09          0.10          0.10
```

Variables with Low Response Rates

Variables with low response rates severely limit the inferences that can be drawn, and usually cannot be used in more complex analyses without resorting to techniques like multiple imputation. The following list reports variable names with more than 300 missing observations, out of 1307 observations total, and reports the number of missing observations for these variables.

In some cases, the variables are follow-up questions where we would expect a low response rate (e.g., smoke), but in other cases there were simply a large number of “I don’t know” responses or outright refusals.

```
##      nojob_months          language          nhoodint
##              1219              1217              1217
##              q68r          years_us              q62r
```

```
##                1214               1146               1145
##       job_pt_choice               cigs         smoke_quit
##                1134               1090               1085
##                q37r     childcare_cost   pub_assist_type1
##                1085               1061               1060
##      childcare_qual       children_act         school_act
##                1050               1047               1046
##     childcare_avail               safe              ctres
##                1044               1017               1017
##             buscell    children_school             q47key
##                1017                969                967
##                 zip            nhoodr2              nhood
##                 866                808                807
##              nhoodr              smoke             job_ft
##                 807                690                688
##          volunteer2 pubtrans_dist_time      pubtrans_dist
##                 626                603                571
##    pubtrans_streets pubtrans_sidewalks    pubtrans_safety
##                 570                567                565
##      pubtrans_rides             q18key               q16r
##                 559                552                552
##                q17r       family_meals               q53r
##                 552                439                430
##        lowinc_progs               q4jr       elderly_supp
##                 396                396                324
##                q4kr
##                 324
```

List of Variable Name Recodings

The following table reports the original and recoded variable names for reference.

original	recoded
state	state
zip	zip
county	county
paq	paq
safe	safe
int51	int51
ctres	ctres
int54	int54
town	town
int57	int57
buscell	buscell
int55	int55
q1	sat_area
q2	area_change
q3	sat_change
q4key	q4key
q4a	healthcare_qual
q4b	locgov_resp
q4c	pubschool_qual
q4d	pubfac_cond

q4e	store_avail
q4f	entertain_avail
q4g	police
q4h	affordable
q4i	employment
q4j	lowinc_progs
q4k	elderly_supp
q4l	children_raise
q5	volunteer
q6	volunteer2
q7	problem_fix
q8	locgov_infl
q9	donated
q10	voter
q11	arts
q12	address_time
q13	housing_type
q14key	q14key
q14a	store_dist
q14b	walk_safety
q14c	bicycle_safety
q14d	pubtrans_sat
q14e	rec_avail
q14f	unsafe
q14g	neighbors_helpful
q14h	neighbors_trust
byr1	byr1
byr2	byr2
age	age
ager	ager
agesny	agesny
q16	pubtrans_rides
q17	pubtrans_dist_time
q18key	q18key
q18a	pubtrans_streets
q18b	pubtrans_sidewalks
q18c	pubtrans_dist
q18d	pubtrans_safety
q19	car_access
q20key	q20key
q20a	org_fire
q20b	org_school
q20c	org_traffic
q21	health
q22key	q22key
q22a	high_bp
q22b	high_cholest
q22c	diabetes
q22d	heart_disease
q22e	asthma
q23	weight_lbs
q24	height_in
q25	health_ins
q26	health_ins_type

q27	has_dr
q28	health_visit
q29	hosp_emerg
q30	dentist_recent
q31	help_avail
q32	mood
q33	exercise
q34	food_money
q35	smoked
q36	smoke
q37	cigs
q38	smoke_quit
q39	fin_manage
q40	future_econ
q41	fin_parents
q42	opp_children
q43key	q43key
q43a	health_delay
q43b	job_loss
q43c	bill_late
q43d	housing_money
q44	noinc_months
q45	marital
q46	num_children
q47key	q47key
q47a	childcare_cost
q47b	childcare_avail
q47c	childcare_qual
q48	children_school
q49	school_act
q50	children_act
q51	role_models
q52	num_adults
q53	family_meals
q54	info_source
q55	computer
llcell	llcell
phonetyp	phonetyp
hisp	hisp
race_m1	race
race_m2	race_m2
race_m3	race_m3
race_m4	race_m4
race_m5	race_m5
racer	racer
q59	english
q60	language
q61	us_born
q62	years_us
educ	educ
income	income
q65	pub_assist
q66_m1	pub_assist_type1
q66_m2	pub_assist_type2

q66_m3	pub_assist_type3
q66_m4	pub_assist_type4
q66_m5	pub_assist_type5
q66_m6	pub_assist_type6
q66_m7	pub_assist_type7
q67	job
q68	nojob_months
q69	job_ft
q70	job_pt_choice
q71	job_sat
nhood	nhood
nhoodint	nhoodint
gender	gender
ring	ring
racew	racew
weight	weight
q16r	q16r
q17r	q17r
q29r	q29r
q33r	q33r
q46r	q46r
q53r	q53r
BMI	bmi
BMIr	bmir
q37r	q37r
q52r	q52r
q62r	q62r
q68r	q68r
nhoodr	nhoodr
nhoodr2	nhoodr2
q4ar	q4ar
q4br	q4br
q4cr	q4cr
q4dr	q4dr
q4er	q4er
q4fr	q4fr
q4gr	q4gr
q4hr	q4hr
q4ir	q4ir
q4jr	q4jr
q4kr	q4kr
q4lr	q4lr
q5r	q5r
q6r	q6r
q7r	q7r
q8r	q8r
q9r	q9r
q10r	q10r
CEscore	cescore