Zhaoxing(Bella) Wu
Final Project

Different Methods of Predicting the Mean of Individual Interests, Dividend, and Rental Income

Abstract. The purpose of this study is to compare and contrast the results of using different trees, forests, or logistic regression to estimate the mean of individual interests, dividend, and rental income with sampling weight. With different classification and regression methods, either the probability of response or the missing value for the variable of interest could be predicted for the estimation of the mean. The result of the study shows that GUIDE (Generalized, Unbiased, Interaction Detection and Estimation) classification forest and GUIDE regression stepwise linear tree give the best estimate among all other methods.

1. Introduction

Nowadays, the majority of people choose to build wealth by investing in stocks, bonds or real estate to ensure consistent returns. Therefore, increasingly more people choose to make earnings not only from regular wages, but also from individual interests, dividend, and rental income as well. With the purpose of analyzing citizens' financial situation, it would be worthwhile to estimate the personal interests, dividend, and rental income, while American Community Survey (ACS) provides a way of doing this. ACS collects Public Use Microdata Sample (PUMS) every year, which includes both person record and household record. In the person record, each observation represents a single person and the variable of interest INTP (the interest, dividend, and net rental income of each person in the past 12 months) is also collected. In this study, the dataset of the person record with 56670 observations and 288 variables is analyzed from the state of Minnesota, which has a population of 5.64 million by the year of 2019, ranked 22 in the United States. The aim of the paper is to apply different methods in predicting the mean of INTP with the sampling weight using the ACS PUMS dataset of Minnesota.

2. Variable Selection and Data Cleaning

The variable of interest is INTP, which has been rounded and top-coded as well as bottom coded, with the minimum of -1300 and the maximum of 254000, as indicated in *Figure 1*. Since the main goal of this project is to estimate the mean of INTP, it is more reasonable to include people with the ability of making income to avoid underestimation. The child labor laws in Minnesota specify that it is illegal to hire minors under the age of 14, while for children during the age of 14 and 15, they can only work outside the school hours and the working hours are extremely limited. Thus, it is significant not to include children less than 15 years old when fitting a model. To address this issue, according to the ACS PUMS dictionary, if the value of INTP is recorded as "bbbbbb", the observations are less than 15 years old and should be deleted. As a result, 9977 observations were removed in total. The allocation flag variable of INTP is FINTP and a value of 0 implies that the INTP of the corresponding observation originally exists, while a value of 1 implies that it is originally missing but imputed by the Census Bureau. If INTP is originally missing, it should be converted as "NA" regardless of imputed value and overall, 7037 observations were converted.

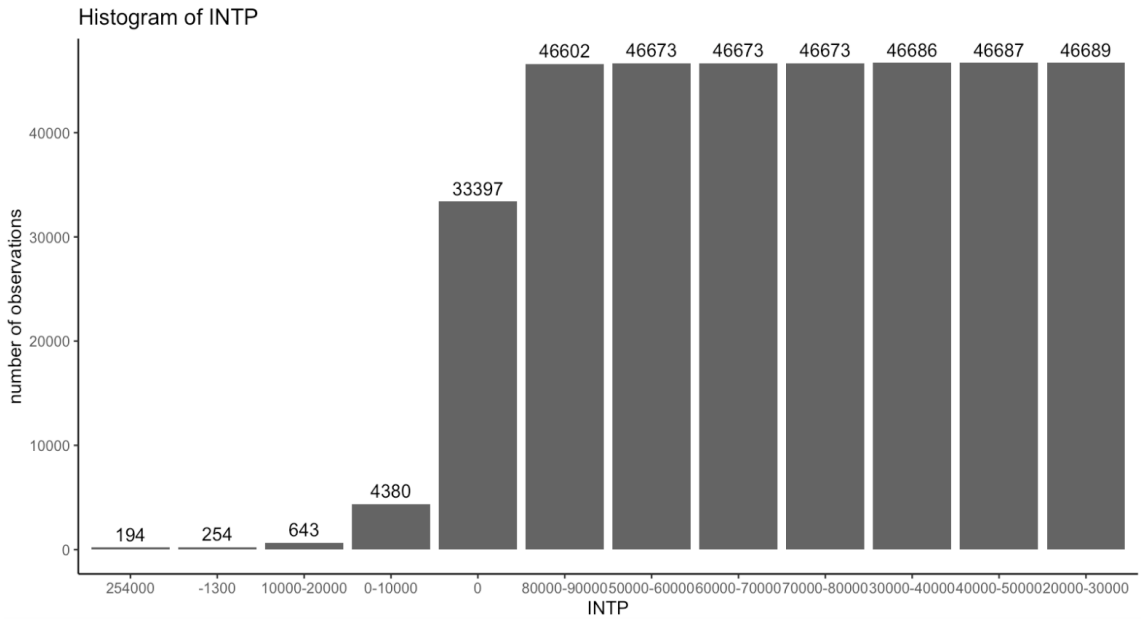


Figure 1: The histogram of INTP (x-axis gives the range of INTP)

There are 288 variables in total in the raw dataset, and however, not every variable is necessary to be included when fitting models. RT, DIVISION, REGION, ST, ADJINC are constant variables that cannot assist the process of predicting missing values or estimating probability of response and should be excluded when fitting a model. SERIALNO and SPORDER are variables that are used to uniquely identify observations with no predicting power, which should be omitted as well. All of the person's weight replicate variables (PWGTP1-80) are for constructing confidence interval which in this case can be ignored. All of the allocation flag variables can be deleted except the flag variable for INTP, the dependent variable. This is because it is necessary to use FINTP to do estimation of probability of response as well as getting more information about INTP to improve the accuracy of prediction, but all the other flag variables might not notably increase the accuracy other than inducing unnecessary computational difficulty, even though they inform whether one observation is originally observed or artificially created.

Categorical variables with too many levels could be time consuming when fitting models and induce potential challenges. For example, "partykit" packages cannot handle categorical variables of more than 31 levels but leading to an integer overrun. The solutions to the problem of high-level categorical variables include removing those variables or combining similar levels into groups. Variables like ANC1P, ANC2P, FOD1P, FOD2P, INDP, LANP, NAICSP, POBP, POWSP, RAC2P, RAC3P, MIGSP, SOCP have more than 40 levels and some even contain more than 400 levels. If treating them as categorical variables, they must be excluded in the model due to the issue of integer overrun, but albeit treating them as numeric variables could lead to biased parameter estimates since they represent different groups. Therefore, those 13 variables must be deleted before fitting models. The second approach is to reduce the number of levels by combining. The variable of occupational recode for 2018 (OCCP) is have 520 levels but many of the different levels could be grouped into subgroups. For example, Lodging Managers, Natural Sciences Managers, Emergency Management Directors can be combined into

a new level “Managers”. By grouping some occupational recodes together into subgroups, OCCP is reduced from 520 levels to only 25 levels and can be included to improve efficiency in predictive model without causing errors.

Correlation and variable dependency might lead to problem of under-performance of some variable selections and those variables should be under particular control. For example, there are two variables both represent the date of entry, DECADE and YOEP, except the difference of the former being the decade of entry and the latter being the year of entry. By deleting DECADE and keeping YOEP, it can provide more detailed information about entry and avoid the problem of correlation. What’s more, DRIVESP is number of vehicles calculated from JWRI, so DRIVESP is linearly dependent on JWRI and only one of the variables can be used to build a model, which in the rest of the paper, JWRI is kept.

Determining the data type of all the variables in the dataset is a paramount step, and the two main data types are qualitative and quantitative. Qualitative variables represent different groups and categories with names or labels such as sex and nativity while quantitative variables are expressed in terms of numbers or a form of measurement. In the ACS PUMS dataset, it seems reasonable to treat variables that indicate person’s weight (PWGTP), age (AGEP), year (CITWP, MARHYP, YOEP), income (INTP, OTP, PAP, RETP, SEMP, SSIP, SSP, WAGP, WKHP, PERNP, PINCP), time (JWMNP, WKMN, JWAP, JWDP), income-to-poverty ratio (POVPIP), vehicle occupancy (JWPIP), and number of race groups (RACNUM) as numeric variables. By treating the other remaining variables as categorical variables, there are 23 numeric variables, 84 categorical variables, 181 excluded variables in total.

After selecting variables and removing certain observations, the raw dataset with 56670 observations and 288 variables is transformed into a new dataset with 46693 observations and 107 variables. All of the selected observations contain missing values. *Table 1* below shows that there are 52 variables with no missing values, 15 variables with fewer than 20% missing values, 16 variables with around or fewer than 50% missing values, and 24 variables with almost completely missing values.

Table 1: Variables in the dataset after cleaning and the number of missing variables

Variable	# missing	Variable	# missing	Variable	# missing	Variable	# missing
GCM	46490	POWPUMA	18507	CIT	0	QTRBIR	0
SFN	46167	WKHP	15014	DDRS	0	RAC1P	0
SFR	46167	WKWN	15014	DEAR	0	RACAIAN	0
GCR	46005	MARHD	12312	DEYE	0	RACASN	0
DRAT	45450	MARHM	12312	DOUT	0	RACBLK	0
CITWP	45286	MARHT	12312	DPHY	0	RACNH	0
ESP	44700	MARHW	12312	DREM	0	RACPI	0
NOP	44700	MARHYP	12312	HIMRKS	0	RACSOR	0
YOEP	44087	COW	11089	HINS1-7	0	RACWHT	0
ENG	43679	OCCP	11089	LANX	0	WAOB	0
MLPA	43139	GCL	8952	MAR	0	FINTP	0
MLPB	43139	INTP	7037	MIG	0	PWGTP	0
MLPCD	43139	WRK	4730	RELSHIPP	0	AGEP	0
MLPE	43139	OC	2465	SCH	0	OIP	0

MLPFG	43139	RC	2465	SCHL	0	PAP	0
MLPH	43139	POVPIP	1752	SEX	0	RETP	0
MLPI	43139	MIL	1467	ANC	0	SEMP	0
MLPJ	43139	NWAB	779	DIS	0	SSIP	0
MLPK	43139	NWAV	779	HICOV	0	SSP	0
VPS	43139	NWLA	779	HISP	0	WAGP	0
DRATX	42206	NWLK	779	MSP	0	PINCP	0
MIGPUMA	41446	NWRE	779	NATIVITY	0	RACNUM	0
SCHG	41254	WKL	779	PRIVCOV	0		
FER	35738	ESR	779	PUBCOV	0		
PAOC	24936	PERNP	779				
JWRIP	22472						
JWMNP	20605						
JWAP	20605						
JWDP	20605						
JWTRNS	18507						

3. Methods

There are two different methods to calculate the estimation of mean. As shown in *Equation 1.1,* one method uses the estimated probability of response $\hat{\pi}$, which can be predicted by setting FINTP as the dependent variable when fitting models with classification trees and forests such as GUIDE, RPART, CTREE, and logistic regression. *S* is the set of non-missing values and *w* is the sampling weight, which in the ACS dataset is represented by PWGTP. As shown in *Equation 1.2,* another method uses the imputed missing value $\hat{y}$, which can be predicted by setting INTP as the dependent variable when fitting models with regression trees and forests such as GUIDE, RPART, CTREE, random forest and CFOREST. $\bar{S}$ is the set of missing values, the complement set of *S*. In this paper, different models are fitted to predict values required in *Equation 1.1* and *Equation 1.2* to estimate the mean of INTP.

$$\mu_{\hat{\pi}} = \left(\sum_{i\in S} \frac{w_i}{\hat{\pi}_i}\right)^{-1} \sum_{i\in S} \frac{w_i y_i}{\hat{\pi}_i} \quad [1]$$

Equation 1.1: The Inverse Probability Weighted (IPW) Estimate of μ Using π̂ (estimated probability of response)

$$\mu_{\hat{y}} = \left(\sum_i w_i\right)^{-1} \left(\sum_{i\in S} w_i y_i + \sum_{j\in \bar{S}} w_j \hat{y}_j\right) \quad [2]$$

Equation 1.2: The Estimate of μ Using ŷ (imputed value of y)

GUIDE stands for "Generalized, Unbiased, Interaction Detection and Estimation" and can be applied to construct tree and forest as well as calculate importance scoring. When applying GUIDE to build classification tree and forest to compute $\hat{\pi}$ or regression tree and forest to compute $\hat{y}$, it is necessary to recode the values in FINTP, the associated missing-value flag variable of INTP since GUIDE can analyze data with more than one missing value code. If FINTP of the observation is originally 0, recode it to be "D" as the valid data value. If FINTP of the observation is originally 1, recode it to be "C" as the missing data value. If INTP of the observation is originally 254000, recode it to be "T" as the top-coded data value. When creating

the data description file, set FINTP or INTP as the dependent variable and assign all other variables to be either categorical, numeric or excluded as mentioned above, except for PWGTP, setting it as a weight variable instead of a numeric variable as indicated in the Appendix. In order to save time and speed up the process of calculating the importance scores, it is desirable to include only important variables that contribute most to prediction by using the facility of importance scoring in GUIDE with a sample of 6000 observations. Using a sample to run importance score might produce fewer importance variables than using the entire dataset, but since the sample is more than 10 percent of the original data, this is not a serious problem. In terms of build trees and forests, unimportant variables are unlikely to be used as predictors in the model, so including only important variables when building models will not affect the results. As shown in *Figure 2.1*, the top three important variables are PINCP, AGEP, HINS3 for least squared regression tree, and 42 important variables should be used to build regression tree and forest. As shown in *Figure 2.2*, the top three important variables are INTP, WRK, ANC for classification tree, and 22 important variables should be used to build classification tree and forest. Even though a sample is used for importance scoring, the entire dataset should be used for tree and forest models to ensure the accuracy.

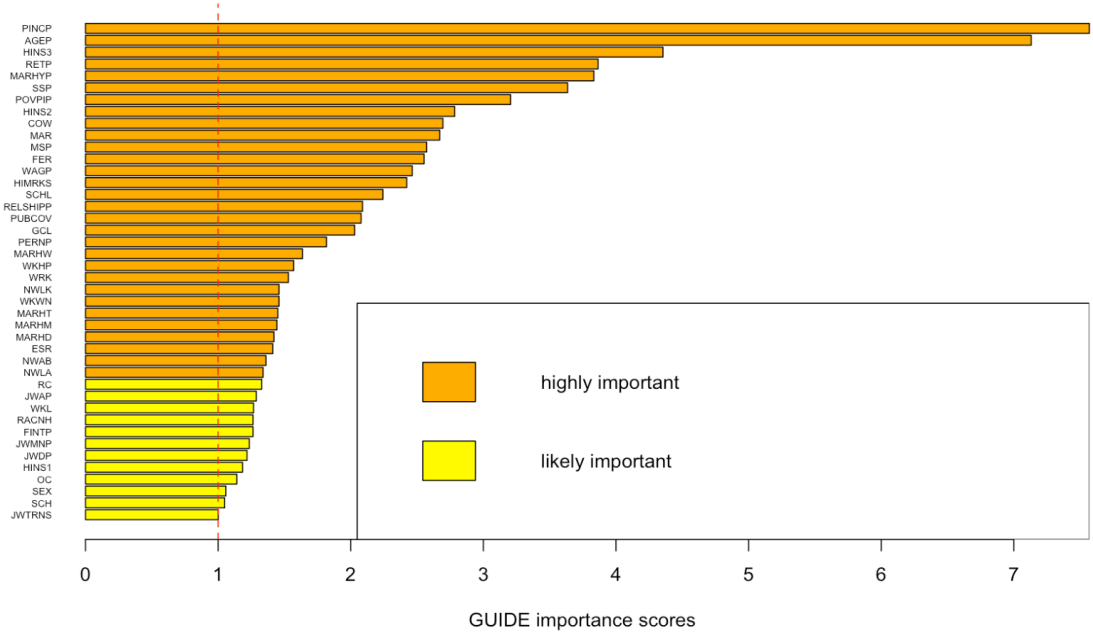


Figure 2.1: GUIDE Importance Score of Least Squared Regression Tree

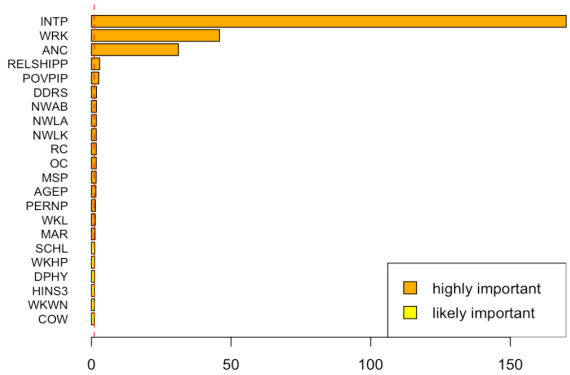


Figure 2.2: GUIDE Importance Score of Classification Tress

RPART and CTREE are methods in “rpart” and “party” packages respectively, used to construct regression and classification trees. No particular modification of the dataset is required to fit these two models. For RPART, the entire dataset after cleaning up is applied to fit regression and classification trees with PWGTP as sampling weight. For CTREE, no missing values are allowed in the dependent variable, so when fitting the regression model, 7307 NA value in INTP should be dropped. Since there is no missing value in FINTP, the classification CTREE can be constructed with the entire dataset after cleaning up. One difference between RPART and CTREE is that the former treats “weight” as sampling weight, while the latter regards “weight” as case weight, and therefore the weight argument is not used in CTREE. From the same package which CTREE belongs to, CFOREST is a method for constructing forest. Similar to CTREE, it cannot take sampling weight as an argument and does not allow missing value in the response variable. Due to the consideration of the time it takes to run CFOREST, only the variables with importance score greater than 2.4 are included in the model, which are 14 variables in total.

Some of the methods do not allow missing values in the dataset, including logistic regression and random forest. There are several ways to impute the missing values in the dataset, and in this paper, maximum likelihood like “Amelia” and sequential regression like “MICE” are used. To save time while imputing, in the case of using FINTP as the dependent variable, only 16 highly important variables are chosen as shown in *Figure 2.1*; in the case of using INTP as the dependent variable, 14 variables with importance score greater than 2.4 are chosen as shown in *Figure 2.2*. And the missing values of INTP are not imputed to avoid cheating, so only 39656 observations are imputed by Amelia and MICE. Since Amelia can only handle categorical data with fewer than 10 levels, RELSHIPP with 19 categories has to be removed. Moreover, perfectly collinear variables would stop the execution of both Amelia and MICE and they should be dropped as well. The variable MSP (married, spouse present/absent) with levels 3, 4, 5, 6 is perfectly collinear with MAR (marital status) with levels 2, 3, 4, 5 respectively, both corresponding to categories of “widowed”, “divorced”, “separated”, “never married” and as a result, the variable MAR is deleted to ensure the smooth execution of Amelia and MICE. Since people who directly purchase insurance from an insurance company are not eligible for subsidy, the variable HIMRKS (subsidized marketplace coverage) of level 0 is perfectly collinear with HINS2 (insurance purchased directly from an insurance company) of level 2, and thus HINS2 is excluded when calling MICE and Amelia. With the imputed results from MICE and Amelia, random forest can be applied to predict the missing values INTP with 12 selected variables, and logistic regression can be used to calculate the probability of response with 14 selected variables, because the dependent variable in logistic regression can only be binary. Since both MICE and Amelia would each produce at least 5 imputed datasets, fitting logistic regression and random forest 5 times and then compute the average of the 5 estimations can avoid the problem of randomness to some degree. Due to the restriction of the computer memory, samples of 30000 observations from the imputed dataset of MICE and Amelia are used in random forest to prevent the problem of “exhausted memory”.

4. Result

Before reporting the estimated mean of INTP by using *Equation 1.1* and *Equation 1.2*, simply calculating the mean of INTP with the sampling weight by ignoring the missing values can assist the comparison of the results from different methods. By using *Equation 2* as shown

below, the weighted mean of INTP is 2007.63. *Table 2.1* and *Table 2.2* gives the estimated mean of each method as well as the corresponding time to fit one dataset.

$$\mu_{weighted} = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}$$

Equation 2: Weighted average (ignoring the missing values in X)

The estimated means by applying *Equation 1.1* are shown below in *Table 2.1*. GUIDE forest produces the largest estimated mean, and the longest time to construct the model. Since GUIDE needs to perform ten-fold cross validation when fitting the model, it is reasonable to have the longest time when running GUIDE. The classification tree with INTP gives the same result as the simple method ignoring the missing value, which is much larger than the other two GUIDE trees without including INTP in the models. Since 231 of the estimated probabilities of response from CTREE are 0, it causes the problem of the denominator of Equation 1.1 being 0. By replacing all the zeros with half of the smallest nonzero estimated probability, the mean is calculated to be 1832.72. The two logistic regressions also have means lower than the method ignoring the missing value, but the logistic regression takes the shortest time compared to all other methods. The smallest estimated mean is RPART, which takes around 10 seconds to fit one dataset.

Table 2.1: The estimated mean of INTP by using different models calculated with Equation 1.1 and the corresponding time

Method	Estimated mean	Time (sec)
GUIDE forest	2047.88	937.6
Simple method ignoring the missing value	2007.62	-
GUIDE tree (with INTP), default univ. splits	2007.62	255.8
CTREE	1987.18*	27.3
Logistic regression, imputed dataset by MICE	1976.92	0.2
GUIDE tree, default univ. splits	1972.78	295.42
Logistic regression, imputed dataset by Amelia	1955.89	0.3
GUIDE tree, 2nd best univ. split at root node	1944.38	291.31
RPART	1928.35	9.6

**replace estimated probability of response of 0 with half of the smallest nonzero estimated probability, namely replacing 0 with 0.007494647*

The estimated means by applying *Equation 1.2* are shown below in *Table 2.2*. GUIDE stepwise linear tree produces the largest estimated mean, followed by CTREE. When applying random forest on imputed data from MICE and Amelia, they both give similar results of around 2090. GUIDE forest gives the result almost the same as the simple method ignoring the missing value. Comparably, the means of GUIDE piecewise constant trees at the default and the 2nd best split are smaller than the simple method. The regression trees constructed by RPART and CTREE take the shortest amount of time. The time of building random forest is as fast as GUIDE piecewise linear trees. It takes more than two hours for constructing CFOREST.

Table 2.2: The estimated mean of INTP by using different models calculated with Equation 1.2 and the corresponding time

Method	Estimated mean	Time (sec)
GUIDE stepwise, default univ. splits	2157.97	1177.2
CTREE	2111.24	33.3
Random forest (MICE)	2090.39	380.7
Random forest (Amelia)	2088.15	254.1
CFOREST	2063.83	8395.7
RPART	2040.04	3.1
Simple method ignoring the missing value	2007.62	-
GUIDE forest	2007.25	1670.5
GUIDE constant, default univ. splits	2000.99	320.1
GUIDE constant, 2nd best univ. split at root node	1966.04	330.2

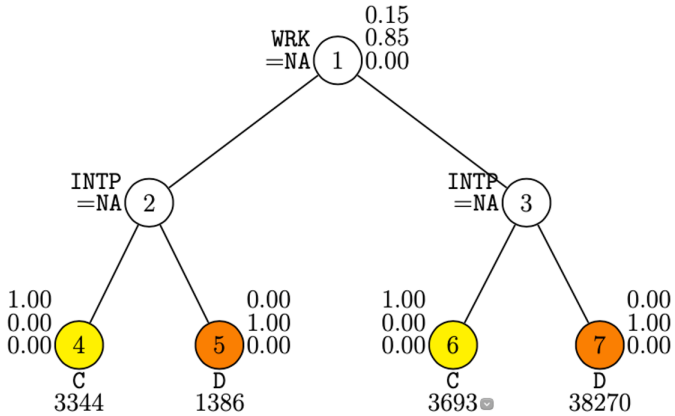
5. Discussion

As mentioned in the last section, the estimation of mean of INTP varies distinctively, ranging from 1928.04 to 2157.97, so it is paramount to study the details of all the models to find out the reasons why, and compare and contrast the accuracy of each method. In order to assess the accuracy, mean squared error and proportion of variance explained could help.

Generally speaking, according to *Table 2.1* and *Table 2.2*, most of the estimations of mean by predicting the probability of response with classification and logistic regression are smaller than the estimate by imputing the missing values. Let's first examine all the classification methods. According to *Figure 3.b*, the GUIDE tree splits at INTP, and the observation is missing if INTP is NA, which is conceptually straightforward to understand. If the tree constructed by GUIDE splits at the second best variable WRK (worked last week) as indicated in *Figure 3.a*, the resulting two subtrees both split on INTP, giving a sense of forcing the tree to split on WRK. These two tree models produce similar estimates, and both have the preference to split on INTP, even though the one at the 2nd best split have 2 more terminal nodes than the tree at the default split. Different from the GUIDE classification tree, both models from RPART and CTREE split on the root of ANC (ancestry recode) with levels $\{1, 2, 3\}$, and then split on NWLK (looking for work) with levels $\{1, 2\}$ and ESR (employment status recode) with levels $\{1, 2, 4\}$, even though these three variables have 135 splits ($\prod(2^{n-1} - 1)$). Compared with the RPART model which has 7 terminal nodes in *Figure 4.a*, the CTREE model involves 76 terminal nodes, making it difficult to interpret. The large CTREE model is mostly due to the large dataset used, since it applies p-value to inspect and choose the best predictor and split. The detailed text of CTREE model could be found in the Appendix. As shown in *Table 2.1*, GUIDE gives mean larger than both RPART and CTREE, which could be explained by the algorithms of dealing with missing values. RPART and CTREE split the missing values by using surrogate variables, which are alternative variables as a substitution for desired split. In contrast, GUIDE assigns missing categorical values to a "missing" category, and this is the reason why GUIDE gives classification tree split on the missing value of INTP or WRK. This effectively treats the missing value as a useful form of information, rather than a burden. Nonetheless, using INTP to predict the missingness of itself is not as informative as using other variables, so it is crucial to remove INTP when fitting the GUIDE tree to proceed analysis. As indicated in *Figure 4*, after excluding INTP, GUIDE tree at the default split contains two splitting nodes WRK, NWLK,

which also exist in RPART and CTREE. In the case of the 2nd best univ split, the tree is much longer and complicated, and prefers to split on WRK and RC which are used twice. According to *Figure 3* and *Figure 4*, the missing values of WRK are used in all GUIDE models to split, which represent persons who did not report the work last week. Therefore, the predictor probably regards all the observations who did not have a job like students, elderly, or the homeless as INTP missing.

a) GUIDE tree, 2nd best univ split



b) GUIDE tree, default univ split

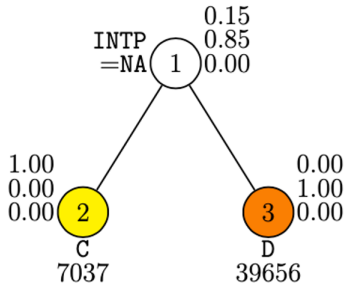


Figure 3: the classification trees build by GUIDE

Generally speaking, according to Table 2, most of the estimations of mean by predicting the probability of response with classification and logistic regression are smaller than the estimate by imputing the missing values. According to *Figure 3.b*, the GUIDE tree splits at INTP, and the observation is missing if INTP is NA, which is conceptually straightforward to understand. If the tree constructed by GUIDE splits at the second best variable WRK, the resulting two subtrees both split on INTP, giving a sense of forcing the tree to split on WRK. These two tree models produce similar estimates, and both have the preference to split on INTP, even though the one at the 2nd best split have 2 more terminal nodes than the tree at the default split. Different from the GUIDE classification tree, both models from RPART and CTREE split on the root of ANC with levels {1, 2, 3}, and then split on NWLK with levels {1, 2} and ESR with levels {1, 2, 4}, even though these three variables have 135 splits ($\prod(2^{n-1}-1)$). Compared with the RPART model which has 7 terminal nodes in *Figure 5.a*, the CTREE model involves 76 terminal nodes, making it difficult to interpret. As shown in *Table 2.1*, GUIDE gives mean larger than both RPART and CTREE, which might be explained by the algorithms of dealing with missing values. RPART and CTREE split the missing values by using surrogate variables, which are alternative variables as a substitution for desired split. In contrast, GUIDE assigns missing categorical values to a "missing" category, and this is the reason why GUIDE gives classification tree split on the missing value of INTP or WRK. This effectively treats the missing value as a useful form of information, rather than a burden. Nonetheless, using INTP to predict the missingness of itself is not as informative as using other variables, so it is crucial to remove INTP when fitting the GUIDE tree to proceed analysis. As indicated in *Figure 4*, after excluding INTP, GUIDE tree at the default split contains two splitting nodes WRK, NWLK, which also exist in RPART and CTREE. In the case of the 2nd best univ split, the tree is much

longer and complicated, and prefers to split on WRK and RC (related child) which are used twice. According to *Figure 3.a* and *Figure 4*, the missing values of WRK are used in GUIDE models to split, and it probably selects all the observations who did not have a job like students, elderly, or the homeless as INTP missing. It seems like employment status is most likely related to the missingness of INTP.

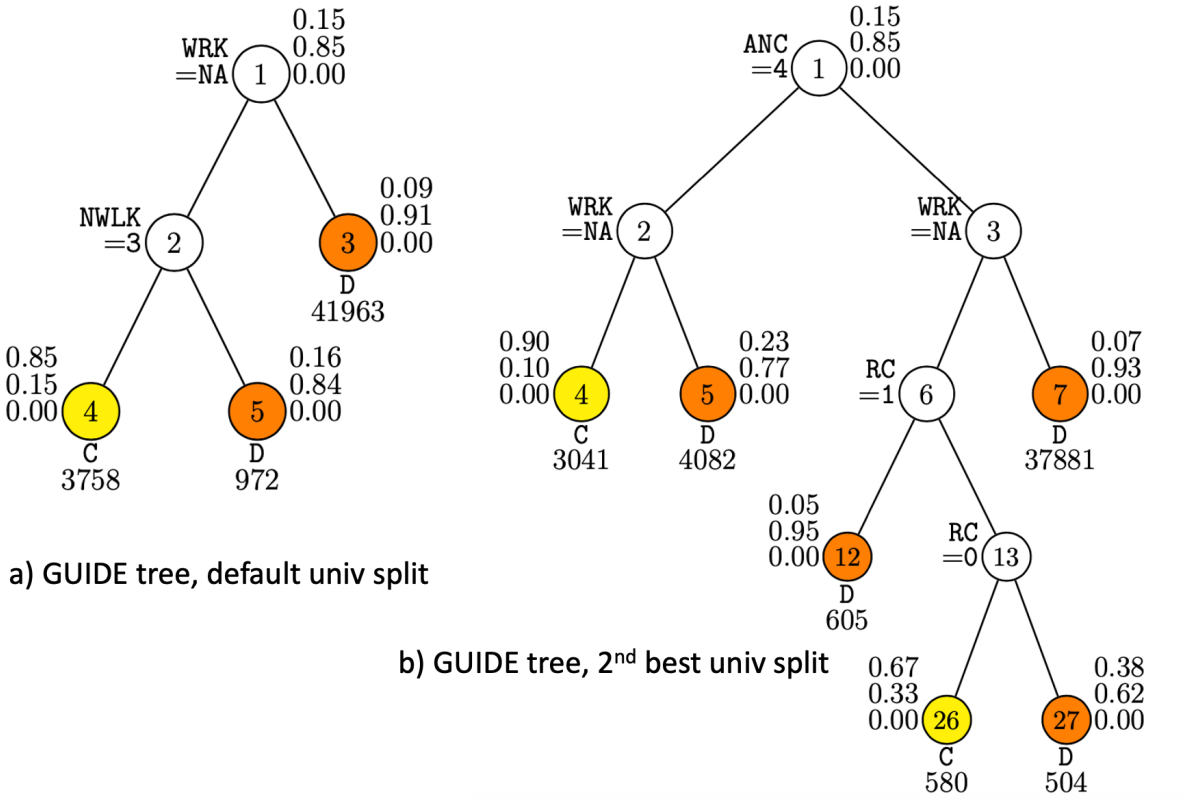


Figure 4: the classification trees build by GUIDE, with INTP excluded

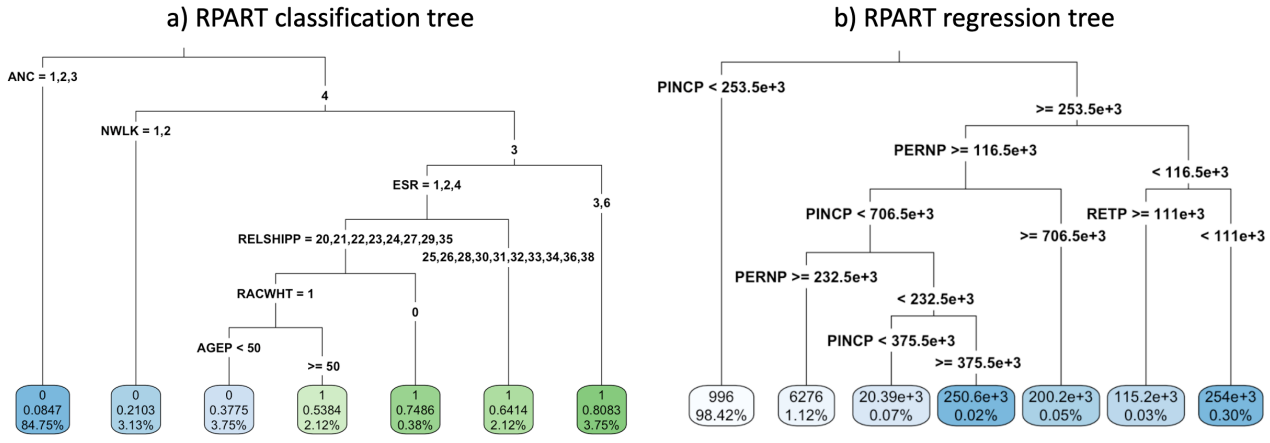


Figure 5: The classification tree and regression tree build by RPART

Logistic regression using data from MICE and Amelia is also constructed to apply *Equation 1.1*. Both results from MICE and Amelia produce a similar model of logistic regression, so only the coefficients of logistic regression from Amelia are shown here. Similar to the GUIDE tree models with INTP excluded, they all regard ANC, NWLK and WRK as statistically significant. What's different, in the logistic regression, RC is not significant, but it is included in GUIDE model in *Figure 4.b*. As the plot indicates, RC is used to classify 1689 observations, however, with a number of observations misclassified, indicating RC might not be a fair predictor. Moreover, originally, there are 2465 observations missing in RC according to Table 1, and thus, the insignificance of RC in the logistic regression might be due to the problem of imputation. Since logistic regression cannot handle missing values, those who are missing in INTP are excluded and the t value of INTP shown below demonstrates that the missingness in INTP is irrelevant to the magnitude of INTP.

Table 3: The estimates, standard error, t value, and p value for coefficients in logistic regression with data from Amelia.

	ESTIMATE	STD. ERROR	T VALUE	PR(>|T|)	
POVPIP	6.402e-06	1.124e-05	0.570	0.568859	
INTP	-5.964e-08	7.877e-08	-0.757	0.448939	
NWLA3	2.018e-02	1.740e-02	1.160	0.246250	
RC1	-1.744e-02	1.317e-02	-1.325	0.185293	
DDRS2	1.284e-02	8.558e-03	1.500	0.133535	
NWLA2	2.921e-02	1.929e-02	1.514	0.129962	
NWAB2	3.365e-02	1.481e-02	2.272	0.023095	*
OC1	3.229e-02	1.344e-02	2.402	0.016298	*
ANC3	2.797e-02	1.069e-02	2.616	0.008891	**
MSP5	5.919e-02	1.805e-02	3.279	0.001041	**
NWLK2	-3.535e-02	1.045e-02	-3.381	0.000723	***
MSP4	1.776e-02	5.206e-03	3.411	0.000648	***
WRK2	3.626e-02	6.639e-03	5.462	4.73e-08	***
NWAB3	9.498e-02	1.662e-02	5.714	1.11e-08	***
NWLK3	8.066e-02	1.254e-02	6.432	1.27e-10	***
(INTERCEPT)	-1.811e-01	2.794e-02	-6.481	9.21e-11	***
MSP3	5.286e-02	6.985e-03	7.569	3.84e-14	***
PERNP	-2.415e-07	3.126e-08	-7.728	1.12e-14	***
WKL2	7.476e-02	7.210e-03	10.368	<2e-16	***
WKL3	7.514e-02	6.286e-03	11.954	<2e-16	***
ANC2	-3.105e-02	3.236e-03	-9.594	<2e-16	***
ANC4	3.856e-01	4.473e-03	86.200	<2e-16	***
MSP2	1.033e-01	1.164e-02	8.874	<2e-16	***
MSP6	4.211e-02	4.483e-03	9.393	<2e-16	***
AGEP	1.745e-03	1.151e-04	15.155	<2e-16	***

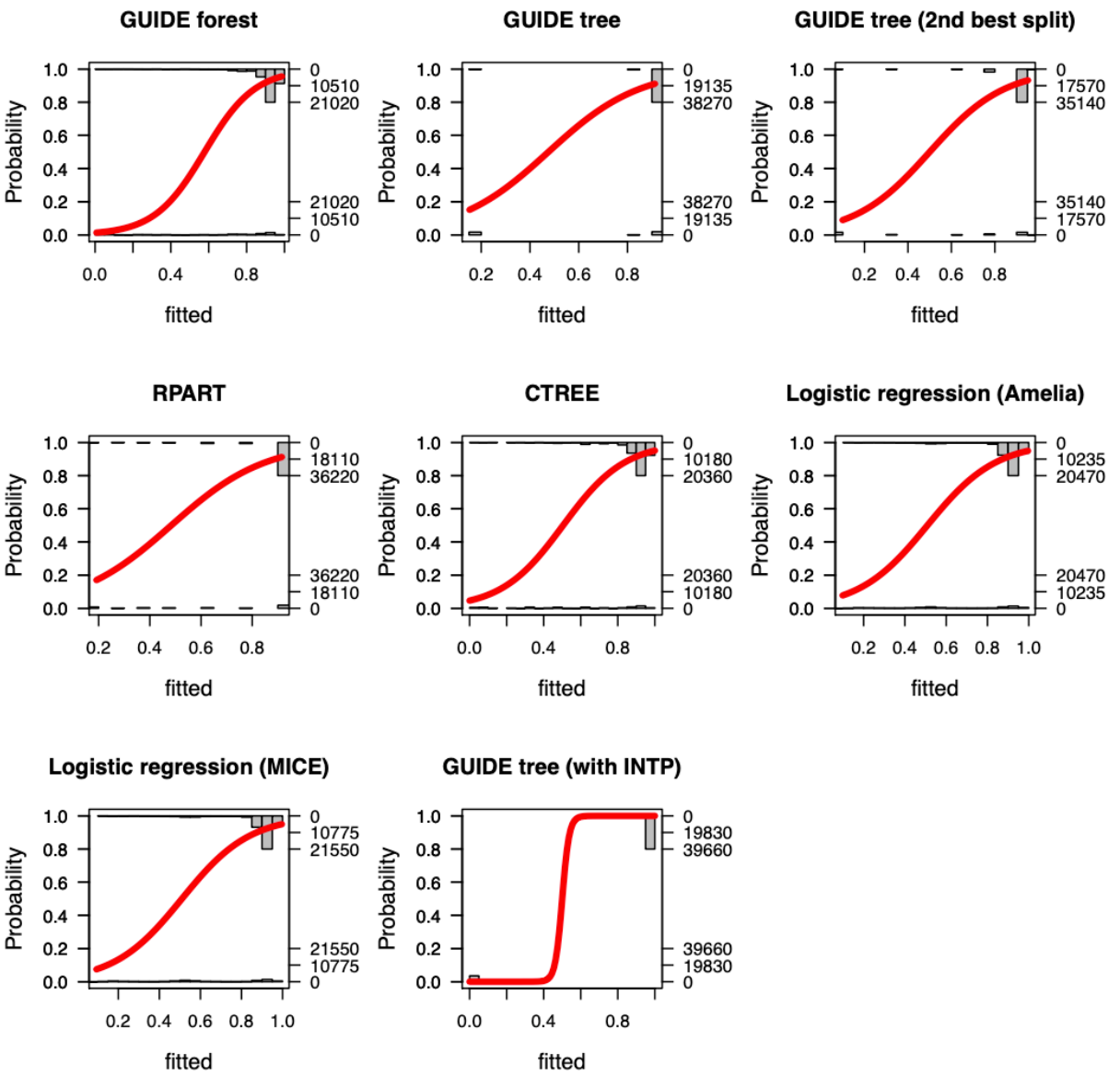


Figure 6: Plots of observed versus fitted values for different classification methods

The plot in *Figure 6* shows the predicted probability versus the actual missingness of INTP, with the x axis as the predicted probability and the right y axis as the number of observations for each bar. The red S shaped curve is the logistic regression fitted to the predicted probability of each model. Unlike all the other models which predict probability of high variance, the GUIDE tree with INTP as the splitting node only returns the predicted probability of either 0.99995 or 0.000018182. *Table 4* provides information about mean squared error, cross validation and number of terminal nodes, and as shown, the mean squared error of the GUIDE tree with INTP as the root is the lowest compared to all the other models, but as mentioned earlier the model is not as informative as the others. In the plots of CTREE, GUIDE forest and the two logistic regressions, there are three pronounce bars at the fitted probability over 0.8, different from other models which only have one remarkable bar at large fitted probability. GUIDE trees at the default split and the second best split and RPART give highly discrete

predicted probability. In the *Table 4*, GUIDE forest is the model that gives the most accurate prediction, while GUIDE tree at the default split is the model with the best model in terms of interpretability without losing much of accuracy. CTREE and RPART have almost the same accuracy, but albeit, CTREE provides a model with much poorer interpretability. In contrast, the logistic regression gives the least accurate prediction. This is most likely due to the nonrandom sample it takes after dropping all the missing value of INTP when building the model, leading to a highly biased estimation.

Table 4: The mean squared error for each classification method, the cross validation for GUIDE models, and number of terminal nodes for tree models

Method	**MSE**	**CV MSE**	**# terminal nodes**
GUIDE tree (with INTP), default univ. splits	2.173085e-09	0.0000428329	2
GUIDE forest	0.07521245	0.0430	-
GUIDE tree, 2nd best univ. split at root node	0.08182362	0.09440387	6
GUIDE tree, default univ. splits	0.08537931	0.09481078	3
CTREE	0.08802209		76
RPART	0.09635999		7
Logistic regression, imputed dataset by Amelia	0.7018822		-
Logistic regression, imputed dataset by MICE	0.7029623		-

Other than using classification in order to predict the probability of response, some of the methods apply regression trees and forests to use *Equation 1.2*. As demonstrated in *Figure 7*, GUIDE piecewise constant trees at the default and the second best split both have the preference to split on PINCP and the only difference between these two models is that the tree at the default split has the root of AGEP. PINCP represents the person's total income, which should include the person's interest, dividend, and rental income and have a positive relationship with INTP. For the GUIDE piecewise constant tree at the default split, the root is AGEP (person's age) less than or equal to 61.5, which classifies people into before retirement and after retirement. For people after retirement, they receive retirement pension monthly which might be lower than the income they make before retirement, so it is reasonable to classify age first and then the total income. The GUIDE stepwise linear tree not only split on PINCP twice and AGEP once similar to GUIDE constant tree, but also split on two other variables PERNP and PUBCOV (public health coverage). PERNP is total person's earning, which different from person's total income, only includes person's wages, bonus or business income but not interest, dividend or rental income. This effectively classifies people who have extreme gap between regular wages and interest income. This tree model is extremely unbalanced according to *Figure 7,* since for the group of people under the age of 18.5, there are no other classifiers for it, probably because most teenagers are unaware of the need to earn interest, dividend, and rental income and specially when some are still continuing their high school education or just starting their college life. In general, GUIDE stepwise tree has more splits, which usually gives a better prediction than GUIDE constant tree. According to *Figure 5.b*, RPART splits on PINCP third and PERNP twice, and use RETP (retirement income) as a splitting node for the group of large value of PINCP and low value of PERNP. In general, models constructed by GUIDE and RPART make prediction by using the people's total income, total earning, and the status of retirement. The regression tree

built by CTREE have 266 terminal nodes, making it hard to interpret, but the root of the tree is still PINCP.

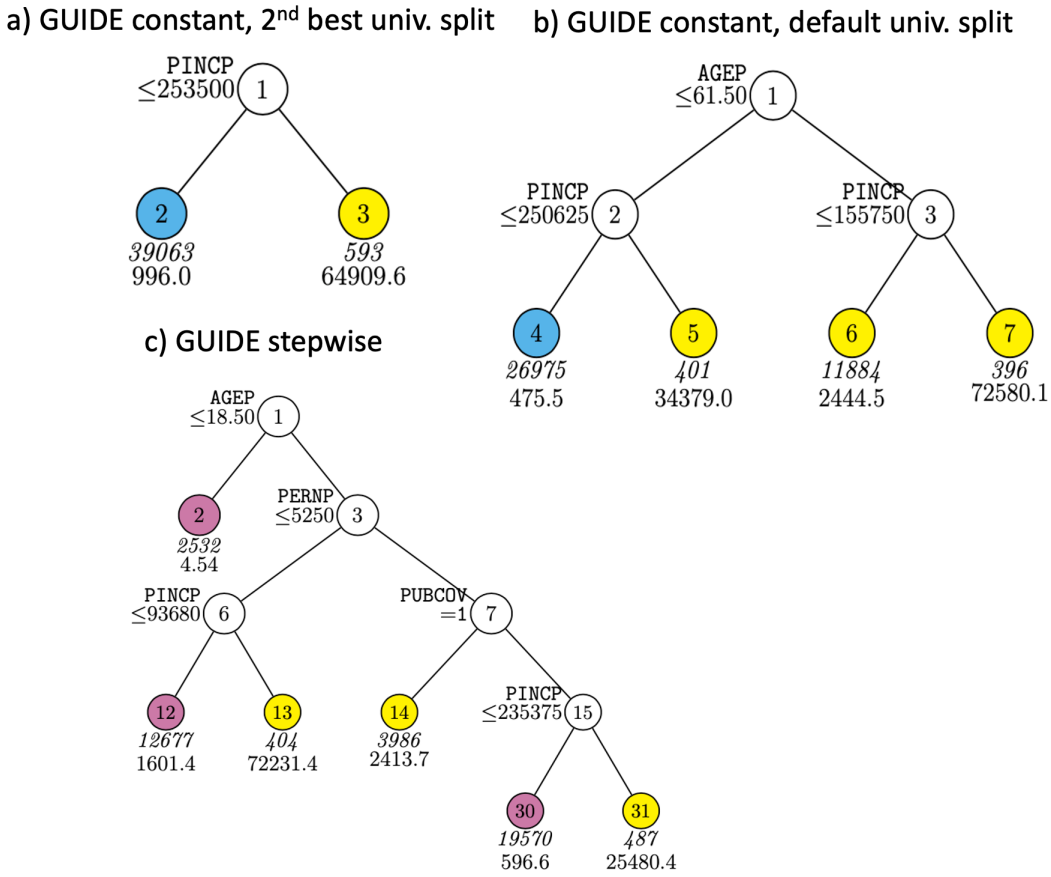


Figure 7: the regression trees built by GUIDE

Figure 8 shows the plot of the observed versus the fitted values of the regression tree and forest models. The four forest models are similar, make great prediction for low value of INTP, but poor prediction for high value of INTP, especially for GUIDE forest which has the most concentrated but deviated predictions for high value of INTP. GUIDE piecewise constant tree can be easily identified by the vertical bands in the plots, and not a lot of the predictions follow the trend of the red line, which might not give as good predictions as other models do. CTREE method seems to make better prediction than the RPART method as there are more points follow the pattern of the red line in the plot of CTREE. However, these two methods have some extremely low or high predictions, when the true value is 254000 or 0 respectively. It is notable that GUIDE stepwise linear model performs well even on high value of INTP, visibly better than all the other methods.

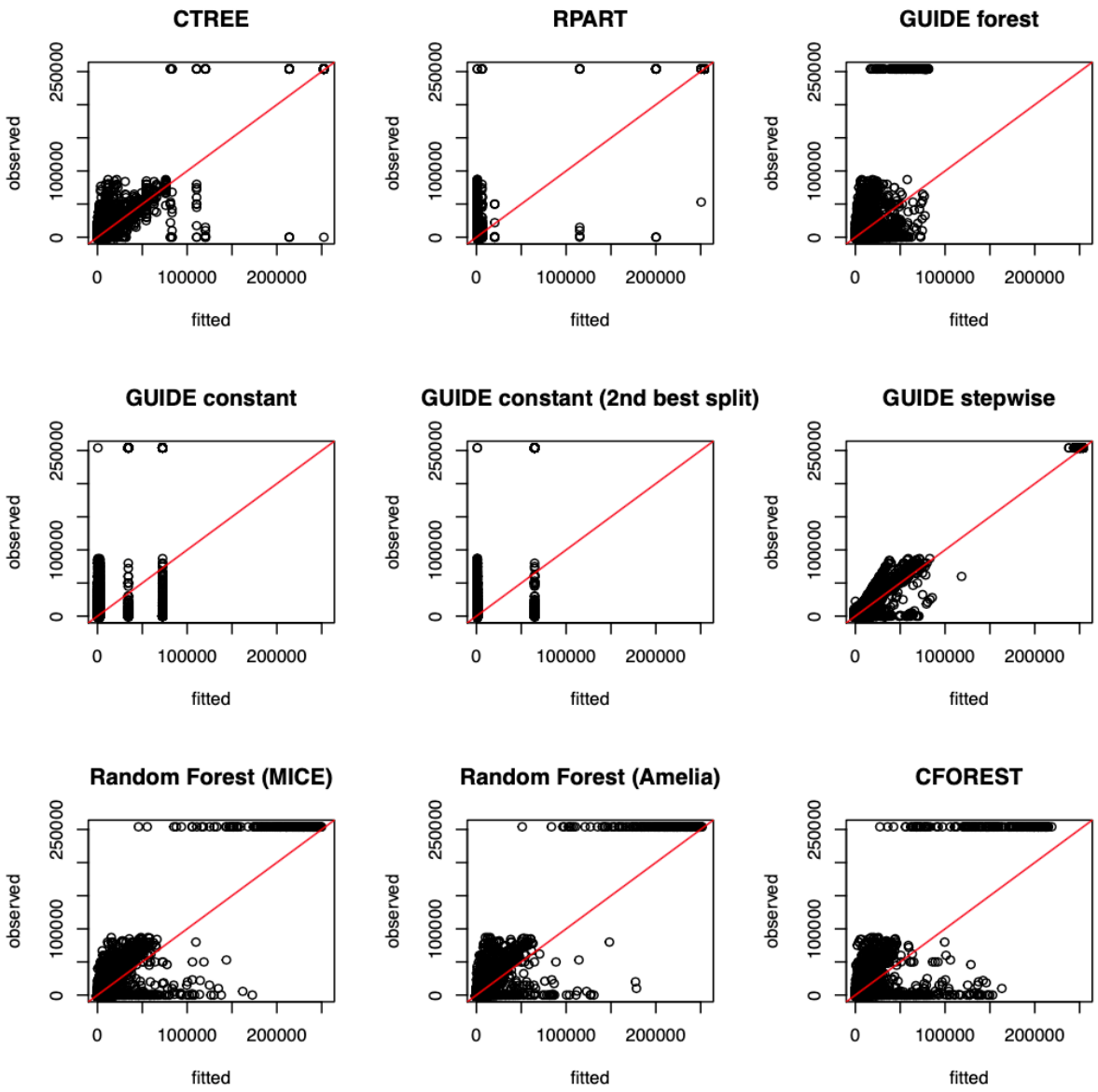


Figure 8: Plots of observed versus fitted values for different regression methods

The table below shows the mean squared error of all regression methods, proportion of variance explained of all GUIDE methods and number of terminal nodes of all trees. GUIDE stepwise linear tree makes the most accurate prediction with great interpretability. Random forest is more accurate than all other forest models. Even though CTREE is as twice as more accurate than RPART, it induces great difficulty in interpreting the tree with 266 terminal nodes. The guide piecewise constant trees make the worse prediction according to the high mean squared error and the low proportion of variance explained. This is probably because the piecewise constant predictions by the tree models only contain only 2 or 3 levels, which estimate the value of INTP ranging from -1300 to 254000 poorly. All in all, the most reasonable prediction of INTP mean is 2157.97 by GUIDE stepwise tree using *Equation 1.2*.

Table 5: The mean squared error for each regression method, the R-squared for GUIDE models, and number of terminal nodes for tree models

Method	MSE	r^2	# terminal nodes
GUIDE stepwise, default univ. splits	16356265	0.9504	6
CTREE	33252264	-	266
Random forest, imputed dataset by Amelia	46597147	-	-
Random forest, imputed dataset by MICE	46937667	-	-
RPART	62480255	-	7
CFOREST	101373930	-	-
GUIDE forest	220507157	0.3748	-
GUIDE constant, 2nd best univ. split at root node	254174219	0.2313	4
GUIDE constant, default univ. splits	261808174	0.2197	2

There are several limitations in this study that is due to the restriction of the dataset, time, and computer memory, and some can be improved in the future. First, the variable of interest INTP is rounded, bottom-coded as well as top-coded, so that every number is estimated to the nearest tens or hundreds, which fails to represent the exact interest, dividend, and rental income of each person. This kind of manipulated data could affect the estimation of INTP to some extent, but not significantly deviating the prediction. The estimation of INTP could be improved if there is a dataset with the actual interest of each person. Second, because of the restriction of time and computer memory, some of the models cannot be constructed with the entire dataset, such as the random forest and logistic regression using MICE and Amelia to impute the dataset. This might not give as accurate prediction as using the entire dataset. Further advancement could be done by using multicore in R to speed up the process when using an entire dataset or run the code on a server. Third, it is hard to compare the accuracy of different models based on one dataset. A fair comparison is using the leave-one-out cross validation, which is hard to achieve with a dataset of over 40000 observations. Even though time and technology restrictions affect the construction and accuracy of different models, generally speaking, the overall trend still holds true and is valuable for future studies.

6. Conclusion

In this study, different forest and tree models are applied to estimate probability of response and missing value of INTP in order to use *Equation 1.1* and *Equation 1.2* to predict the mean. In general, GUIDE regression stepwise tree and GUIDE classification forest gives the best estimate of the mean of interest, dividend and rental income of each person. However, GUIDE piecewise constant trees do not perform well due to the few levels prediction they make, while the logistic regression gives the poorest prediction because of the biased sample it takes to fit the model. In the future study, some improvements could be done in terms of the equipment used to run models as well as more detailed and accurate dataset. Appendix is attached below for references.

Appendix

Input file for classification importance scoring

```
GUIDE       (do not edit this file unless you know what you are doing)
 36.2       (version of GUIDE that generated this file)
 2          (1=model fitting, 2=importance or DIF scoring, 3=data conversion)
"out.txt"  (name of output file)
 1          (1=classification, 2=regression, 3=propensity score grouping)
 1          (1=univariate and interaction splits, 2=skip interactions)
"desc.txt"  (name of data description file)
 1          (1=estimated priors, 2=equal priors, 3=other priors)
 1          (1=unit misclassification costs, 2=other)
 2          (1=split point from quantiles, 2=use exhaustive search)
 1          (1=default max. number of split levels, 2=specify no. in next line)
 1          (1=default min. node size, 2=specify min. value in next line)
 2          (0=no LaTeX code, 1=tree without node numbers, 2=tree with node numbers)
"code.tex" (latex file name)
 1          (1=color terminal nodes, 2=no colors)
 2          (0=#errors, 1=sample sizes, 2=sample proportions, 3=posterior probs, 4=nothing)
 1          (1=no storage, 2=store fit and split variables, 3=store split variables and values)
 2          (1=do not create description file for selected variables, 2=create the file)
 1          (1=exclude non-selected variables, 2=exclude selected variables)
"descclass.txt"  (name of new description file to be created)
 1          (1=create file for importance scores, 2=do not create)
"imp.txt" (file name for importance scores)
 1          (rank of top variable to split root node)
```

Input file for regression importance scoring

```
GUIDE       (do not edit this file unless you know what you are doing)
 36.2       (version of GUIDE that generated this file)
 2          (1=model fitting, 2=importance or DIF scoring, 3=data conversion)
"out.txt"  (name of output file)
 2          (1=classification, 2=regression, 3=propensity score grouping)
 1          (1=linear, 2=quantile, 3=Poisson, 4=censored response, 5=multiresponse or
itemresponse, 6=longitudinal with T vars, 7=logistic)
 1          (1=least squares, 2=least median of squares)
 1          (1=interaction tests, 2=skip them)
"desc.txt"  (name of data description file)
 2          (1=split point from quantiles, 2=use exhaustive search)
 1          (1=default max. number of split levels, 2=specify no. in next line)
 1          (1=default min. node size, 2=specify min. value in next line)
 2          (0=no LaTeX code, 1=tree without node numbers, 2=tree with node numbers)
"code.tex" (latex file name)
 1          (0=all white,1=yellow-skyblue,2=yellow-purple,3=yellow-orange,4=orange-
skyblue,5=yellow-red,6=orange-purple,7=grayscale)
 1          (1=no storage, 2=store fit and split variables, 3=store split variables and values)
 1          (1=do not save, 2=save regressor names in a file)
 2          (1=do not create description file for selected variables, 2=create the file)
 1          (1=exclude non-selected variables, 2=exclude selected variables)
"descreg.txt"  (name of new description file to be created)
 1          (1=create file for importance scores, 2=do not create)
"imp.txt" (file name for importance scores)
 1          (rank of top variable to split root node)
```

Input file for regression tree, piecewise constant, at default split

```
GUIDE       (do not edit this file unless you know what you are doing)
 36.2       (version of GUIDE that generated this file)
 1          (1=model fitting, 2=importance or DIF scoring, 3=data conversion)
"out.txt"  (name of output file)
 1          (1=one tree, 2=ensemble)
 2          (1=classification, 2=regression, 3=propensity score grouping)
 1          (1=linear, 2=quantile, 3=Poisson, 4=censored response, 5=multiresponse or
itemresponse, 6=longitudinal with T vars, 7=logistic)
 1          (1=least squares, 2=least median of squares)
 3          (0=stepwise, 1=multiple linear, 2=simple polynomial, 3=constant, 4=ANCOVA)
 1          (1=interaction tests, 2=skip them)
 1          (0=tree with fixed no. of nodes, 1=prune by CV, 2=no pruning)
"descreg.txt"  (name of data description file)
       10  (number of cross-validations)
 1          (1=mean-based CV tree, 2=median-based CV tree)
    0.500  (SE number for pruning)
 2          (1=unweighted, 2=weighted error estimates during pruning)
 2          (1=split point from quantiles, 2=use exhaustive search)
 1          (1=default max. number of split levels, 2=specify no. in next line)
 1          (1=default min. node size, 2=specify min. value in next line)
 2          (0=no LaTeX code, 1=tree without node numbers, 2=tree with node numbers)
"code.tex" (latex file name)
 1          (0=all white,1=yellow-skyblue,2=yellow-purple,3=yellow-orange,4=orange-
skyblue,5=yellow-red,6=orange-purple,7=grayscale)
 1          (1=no storage, 2=store fit and split variables, 3=store split variables and values)
 1          (1=do not save, 2=save regressor names in a file)
 2          (1=do not save fitted values and node IDs, 2=save in a file)
"fit.txt" (file name for fitted values and node IDs)
 1          (1=do not write R function, 2=write R function)
 1          (rank of top variable to split root node)
```

Input file for regression tree, piecewise constant, at 2nd best split

```
GUIDE       (do not edit this file unless you know what you are doing)
 36.2       (version of GUIDE that generated this file)
 1          (1=model fitting, 2=importance or DIF scoring, 3=data conversion)
"out.txt"  (name of output file)
 1          (1=one tree, 2=ensemble)
 2          (1=classification, 2=regression, 3=propensity score grouping)
 1          (1=linear, 2=quantile, 3=Poisson, 4=censored response, 5=multiresponse or
itemresponse, 6=longitudinal with T vars, 7=logistic)
 1          (1=least squares, 2=least median of squares)
 3          (0=stepwise, 1=multiple linear, 2=simple polynomial, 3=constant, 4=ANCOVA)
 1          (1=interaction tests, 2=skip them)
 1          (0=tree with fixed no. of nodes, 1=prune by CV, 2=no pruning)
"descreg.txt"  (name of data description file)
       10  (number of cross-validations)
 1          (1=mean-based CV tree, 2=median-based CV tree)
    0.500  (SE number for pruning)
 2          (1=unweighted, 2=weighted error estimates during pruning)
 2          (1=split point from quantiles, 2=use exhaustive search)
 1          (1=default max. number of split levels, 2=specify no. in next line)
 1          (1=default min. node size, 2=specify min. value in next line)
 2          (0=no LaTeX code, 1=tree without node numbers, 2=tree with node numbers)
"code.tex" (latex file name)
 1          (0=all white,1=yellow-skyblue,2=yellow-purple,3=yellow-orange,4=orange-
skyblue,5=yellow-red,6=orange-purple,7=grayscale)
 1          (1=no storage, 2=store fit and split variables, 3=store split variables and values)
 1          (1=do not save, 2=save regressor names in a file)
 2          (1=do not save fitted values and node IDs, 2=save in a file)
"fit.txt" (file name for fitted values and node IDs)
 1          (1=do not write R function, 2=write R function)
 2          (rank of top variable to split root node)
```

Input file for regression tree, stepwise linear, at default split

```
GUIDE       (do not edit this file unless you know what you are doing)
 36.2       (version of GUIDE that generated this file)
 1          (1=model fitting, 2=importance or DIF scoring, 3=data conversion)
"out.txt"   (name of output file)
 1          (1=one tree, 2=ensemble)
 2          (1=classification, 2=regression, 3=propensity score grouping)
 1          (1=linear, 2=quantile, 3=Poisson, 4=censored response, 5=multiresponse or
itemresponse, 6=longitudinal with T vars, 7=logistic)
 1          (1=least squares, 2=least median of squares)
 0          (0=stepwise, 1=multiple linear, 2=simple polynomial, 3=constant, 4=ANCOVA)
 1          (1=forward+backward, 2=forward, 3=all subsets)
    0       (max. number of variables to be selected; 0=max. possible)
 4.00       (f-to-enter)
 3.99       (f-to-delete)
 3          (0=no truncation, 1=node range, 2=+10% node range, 3=global range)
 1          (1=interaction tests, 2=skip them)
 1          (0=tree with fixed no. of nodes, 1=prune by CV, 2=no pruning)
"descreg.txt"  (name of data description file)
 2          (missing regressor values: 1=separate models, 2=impute with means, 3=constant model)
       10   (number of cross-validations)
 1          (1=mean-based CV tree, 2=median-based CV tree)
    0.500   (SE number for pruning)
 2          (1=unweighted, 2=weighted error estimates during pruning)
 1          (1=accept default splitting fraction, 2=change it)
 1          (1=default max. number of split levels, 2=specify no. in next line)
 1          (1=default min. node size, 2=specify min. value in next line)
 2          (0=no LaTeX code, 1=tree without node numbers, 2=tree with node numbers)
"code.tex" (latex file name)
 2          (0=all white,1=yellow-skyblue,2=yellow-purple,3=yellow-orange,4=orange-
skyblue,5=yellow-red,6=orange-purple,7=grayscale)
 1          (1=no storage, 2=store fit and split variables, 3=store split variables and values)
 1          (1=do not save, 2=save regressor names in a file)
 2          (1=do not save fitted values and node IDs, 2=save in a file)
"fit.txt"  (file name for fitted values and node IDs)
 1          (1=do not write R function, 2=write R function)
 1          (rank of top variable to split root node)
```

Input file for regression forest

```
GUIDE       (do not edit this file unless you know what you are doing)
 36.2       (version of GUIDE that generated this file)
 1          (1=model fitting, 2=importance or DIF scoring, 3=data conversion)
"out.txt"  (name of output file)
 2          (1=one tree, 2=ensemble)
 2          (1=bagging, 2=rforest)
 2          (1=random splits of missing values, 2=nonrandom)
 2          (1=classification, 2=regression)
 2          (1=interaction tests, 2=skip them)
"descreg.txt"  (name of data description file)
 1          (1=accept default number of trees, 2=change)
 1          (1=accept default number of variables for splitting, 2=change it)
 1          (1=split point from quantiles, 2=use exhaustive search)
 1          (1=accept default splitting fraction, 2=change it)
 1          (1=default max. number of split levels, 2=specify no. in next line)
 1          (1=default min. node size, 2=specify min. value in next line)
"fit.txt" (file name for predicted values)
 1          (rank of top variable to split root node)
```

Input file for classification tree, piecewise constant, at default split

```
GUIDE       (do not edit this file unless you know what you are doing)
 36.2       (version of GUIDE that generated this file)
 1          (1=model fitting, 2=importance or DIF scoring, 3=data conversion)
"out.txt"  (name of output file)
 1          (1=one tree, 2=ensemble)
 1          (1=classification, 2=regression, 3=propensity score grouping)
 1          (1=simple model, 2=nearest-neighbor, 3=kernel)
 1          (0=linear 1st, 1=univariate 1st, 2=skip linear, 3=skip linear and interaction)
 1          (0=tree with fixed no. of nodes, 1=prune by CV, 2=by test sample, 3=no pruning)
"descclass.txt"  (name of data description file)
       10  (number of cross-validations)
 1          (1=mean-based CV tree, 2=median-based CV tree)
    0.500  (SE number for pruning)
 1          (1=estimated priors, 2=equal priors, 3=other priors)
 1          (1=unit misclassification costs, 2=other)
 2          (1=split point from quantiles, 2=use exhaustive search)
 1          (1=default max. number of split levels, 2=specify no. in next line)
 1          (1=default min. node size, 2=specify min. value in next line)
 2          (0=no LaTeX code, 1=tree without node numbers, 2=tree with node numbers)
"code.tex" (latex file name)
 1          (1=color terminal nodes, 2=no colors)
 2          (0=#errors, 1=sample sizes, 2=sample proportions, 3=posterior probs, 4=nothing)
 1          (1=no storage, 2=store fit and split variables, 3=store split variables and values)
 2          (1=do not save fitted values and node IDs, 2=save in a file)
"fit.txt" (file name for fitted values and node IDs)
 1          (1=do not write R function, 2=write R function)
 1          (rank of top variable to split root node)
```

Input file for classification tree, piecewise constant, at 2nd best split

```
GUIDE       (do not edit this file unless you know what you are doing)
 36.2       (version of GUIDE that generated this file)
 1          (1=model fitting, 2=importance or DIF scoring, 3=data conversion)
"out.txt"  (name of output file)
 1          (1=one tree, 2=ensemble)
 1          (1=classification, 2=regression, 3=propensity score grouping)
 1          (1=simple model, 2=nearest-neighbor, 3=kernel)
 1          (0=linear 1st, 1=univariate 1st, 2=skip linear, 3=skip linear and interaction)
 1          (0=tree with fixed no. of nodes, 1=prune by CV, 2=by test sample, 3=no pruning)
"descclass.txt"  (name of data description file)
       10  (number of cross-validations)
 1          (1=mean-based CV tree, 2=median-based CV tree)
    0.500  (SE number for pruning)
 1          (1=estimated priors, 2=equal priors, 3=other priors)
 1          (1=unit misclassification costs, 2=other)
 2          (1=split point from quantiles, 2=use exhaustive search)
 1          (1=default max. number of split levels, 2=specify no. in next line)
 1          (1=default min. node size, 2=specify min. value in next line)
 2          (0=no LaTeX code, 1=tree without node numbers, 2=tree with node numbers)
"code.tex" (latex file name)
 1          (1=color terminal nodes, 2=no colors)
 2          (0=#errors, 1=sample sizes, 2=sample proportions, 3=posterior probs, 4=nothing)
 1          (1=no storage, 2=store fit and split variables, 3=store split variables and values)
 2          (1=do not save fitted values and node IDs, 2=save in a file)
"fit.txt" (file name for fitted values and node IDs)
 1          (1=do not write R function, 2=write R function)
 2          (rank of top variable to split root node)
```

Input file for classification forest

```
GUIDE       (do not edit this file unless you know what you are doing)
 36.2       (version of GUIDE that generated this file)
 1          (1=model fitting, 2=importance or DIF scoring, 3=data conversion)
"out.txt"  (name of output file)
 2          (1=one tree, 2=ensemble)
 2          (1=bagging, 2=rforest)
 2          (1=random splits of missing values, 2=nonrandom)
 1          (1=classification, 2=regression)
 2          (1=interaction tests, 2=skip them)
"descclass.txt"  (name of data description file)
 1          (1=accept default number of trees, 2=change)
 1          (1=accept default number of variables for splitting, 2=change it)
 1          (1=estimated priors, 2=equal priors, 3=other priors)
 1          (1=unit misclassification costs, 2=other)
 1          (1=split point from quantiles, 2=use exhaustive search)
 1          (1=accept default splitting fraction, 2=change it)
 1          (1=default max. number of split levels, 2=specify no. in next line)
 1          (1=default min. node size, 2=specify min. value in next line)
"fit.txt" (file name for predicted class and probability estimates)
 1          (rank of top variable to split root node)
```

Output file of classification forest

Random forest of classification trees
No pruning
Data description file: descclass.txt
Training sample file: data.txt
Missing value code: NA
Records in data file start on line 2
5 N variables changed to S
D variable is FINTP
Number of records in data file: 46693
Length of longest entry in data file: 13
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Number of classes: 3
Training sample class proportions of D variable FINTP:

Class	#Cases	Proportion
C	7037	0.15070781
D	39654	0.84924935
T	2	0.00004283

Summary information for training sample of size 46693
d=dependent, b=split and fit cat variable using indicator variables,
c=split-only categorical, i=fit-only categorical (via indicators),
s=split-only numerical, n=split and fit numerical, f=fit-only numerical,
m=missing-value flag variable, p=periodic variable, w=weight

Column	Name		Minimum	Maximum	#Codes/ Levels/ Periods	#Missing
10	AGEP	s	15.00	95.00		
13	COW	c			9	11089
14	DDRS	c			2	
18	DPHY	c			2	
30	HINS3	c			2	
40	MAR	c			5	
57	NWAB	c			3	779
59	NWLA	c			3	779
60	NWLK	c			3	779
64	RELSHIPP	c			19	
68	SCHL	c			24	
74	WKHP	s	1.000	99.00		15014
75	WKL	c			3	779
76	WKWN	s	1.000	52.00		15014
77	WRK	c			2	4730
79	ANC	c			4	
97	MSP	c			6	

```
   101  OC        c                          2      2465
   104  PERNP     s   -8000.       0.8310E+06              779
   107  POVPIP    s    0.000        501.0                 1752
   124  RC        c                          2      2465
   167  FINTP     d                          3

    Total  #cases w/   #missing
   #cases    miss. D  ord. vals   #X-var   #N-var   #F-var   #S-var
    46693          0      15859      266        0        0        5
   #P-var   #M-var   #B-var   #C-var   #I-var
        0        0        0       16        0
Number of cases used for training: 46693
Number of split variables: 21
Number of cases excluded due to 0 weight or missing D: 0

Number of trees in ensemble: 500
Number of variables used for splitting: 8
Simple node models
Estimated priors
Unit misclassification costs
Univariate split highest priority
No interaction splits
Fraction of cases used for splitting each node: .0021
Maximum number of split levels: 30
Minimum node sample size: 233
Mean number of terminal nodes:    145.5

Classification matrix for training sample:
Predicted       True class
class             C         D         T
C              3045       296         0
D              3992     39358         2
T                 0         0         0
Total          7037     39654         2

Number of cases used for tree construction: 46693
Number misclassified: 4290
Resubstitution estimate of mean misclassification cost: .0919

Number of OOB cases: 46693
Number OOB misclassified: 4296
OOB estimate of mean misclassification cost: .0920
Mean number of trees per OOB observation: 183.92

Predicted class probabilities are stored in fit.txt
Elapsed time in seconds: 937.56
```

Output file of regression forest

Random forest of GUIDE least-squares regression trees
No pruning
Data description file: descreg.txt
Training sample file: data.txt
Missing value code: NA
Records in data file start on line 2
13 N variables changed to S
D variable is INTP
Piecewise constant model
Number of records in data file: 46693
Length of longest entry in data file: 13
Missing values found in D variable
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Smallest and largest positive weights are 1.0000E+00 and 2.4080E+03

Summary information for training sample of size 39656 (excluding observations with non-positive weight or missing values in d, e, t, r or z variables)
d=dependent, b=split and fit cat variable using indicator variables,
c=split-only categorical, i=fit-only categorical (via indicators),
s=split-only numerical, n=split and fit numerical, f=fit-only numerical,
m=missing-value flag variable, p=periodic variable, w=weight

Column	Name		Minimum	Maximum	#Codes/ Levels/ Periods	#Missing
9	PWGTP	w	1.000	2408.		
10	AGEP	s	15.00	95.00		
13	COW	c			9	8906
23	FER	c			2	30173
24	GCL	c			2	7529
27	HIMRKS	c			3	
28	HINS1	c			2	
29	HINS2	c			2	
30	HINS3	c			2	
35	INTP	d	-1300.	0.2540E+06		
36	JWMNP	s	1.000	163.0		17020
38	JWTRNS	c			12	15212
40	MAR	c			5	
41	MARHD	c			2	10258
42	MARHM	c			2	10258
43	MARHT	c			3	10258
44	MARHW	c			2	10258
45	MARHYP	s	1940.	2019.		10258
57	NWAB	c			3	672

```
 59  NWLA       c                          3       672
 60  NWLK       c                          3       672
 64  RELSHIPP   c                         19
 65  RETP       s    0.000       0.1420E+06
 66  SCH        c                          3
 68  SCHL       c                         24
 70  SEX        c                          2
 72  SSP        s    0.000       0.3690E+05
 73  WAGP       s    0.000       0.4760E+06
 74  WKHP       s    1.000       99.00              12232
 75  WKL        c                          3       672
 76  WKWN       s    1.000       52.00              12232
 77  WRK        c                          2      1386
 86  ESR        c                          5       672
 92  JWAP       s    1.000       285.0              17020
 93  JWDP       s    1.000       150.0              17020
 97  MSP        c                          6
101  OC         c                          2      1665
104  PERNP      s   -8000.       0.8310E+06          672
105  PINCP      s   -9300.       0.8590E+06
107  POVPIP     s    0.000       501.0              1126
111  PUBCOV     c                          2
119  RACNH      c                          2
124  RC         c                          2      1665
167  FINTP      c                          3

 Total  #cases w/   #missing
#cases    miss. D  ord. vals   #X-var   #N-var   #F-var   #S-var
 46693       7037      30071      244        0        0       13
#P-var   #M-var   #B-var   #C-var   #I-var
    0        0        0       29        0
Weight variable PWGTP in column: 9
Number of cases used for training: 39656
Number of split variables: 42
Number of cases excluded due to 0 weight or missing D: 7037

Number of trees in ensemble: 500
Number of variables used for splitting: 15
No nodewise interaction tests
Fraction of cases used for splitting each node: .0025
Maximum number of split levels: 30
Minimum node sample size: 198
Mean number of terminal nodes:    147.8
Resubstitution estimate of mean squared error: 220507157.4286
based on number of training cases: 39656
Proportion of variance (R-squared) explained by ensemble model: 0.3748
```

Number of OOB cases: 39656
OOB estimate of mean squared error: 223119640.7202
Mean number of trees per OOB observation: 183.93

Number of test cases with 0 weight and nonmissing responses = 0
Observed and fitted values are stored in fit.txt
Elapsed time in seconds: 1670.5

Output file of classification tree with INTP included

Classification tree
Pruning by cross-validation
Data description file: descclass.txt
Training sample file: data.txt
Missing value code: NA
Records in data file start on line 2
6 N variables changed to S
D variable is FINTP
Number of records in data file: 46693
Length of longest entry in data file: 13
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Number of classes: 3
Training sample class proportions of D variable FINTP:

Class	#Cases	Proportion
C	7037	0.15070781
D	39654	0.84924935
T	2	0.00004283

Summary information for training sample of size 46693
d=dependent, b=split and fit cat variable using indicator variables,
c=split-only categorical, i=fit-only categorical (via indicators),
s=split-only numerical, n=split and fit numerical, f=fit-only numerical,
m=missing-value flag variable, p=periodic variable, w=weight

Column	Name		Minimum	Maximum	#Codes/ Levels/ Periods	#Missing
10	AGEP	s	15.00	95.00		
13	COW	c			9	11089
14	DDRS	c			2	
18	DPHY	c			2	
30	HINS3	c			2	
35	INTP	s	-1300.	0.2540E+06		7037
40	MAR	c			5	
57	NWAB	c			3	779
59	NWLA	c			3	779
60	NWLK	c			3	779
64	RELSHIPP	c			19	
68	SCHL	c			24	
74	WKHP	s	1.000	99.00		15014
75	WKL	c			3	779
76	WKWN	s	1.000	52.00		15014
77	WRK	c			2	4730
79	ANC	c			4	

```
 97  MSP       c                          6
101  OC        c                          2      2465
104  PERNP     s   -8000.       0.8310E+06             779
107  POVPIP    s    0.000        501.0                1752
124  RC        c                          2      2465
167  FINTP     d                          3

 Total  #cases w/   #missing
#cases    miss. D  ord. vals   #X-var   #N-var   #F-var   #S-var
 46693          0      19927      265        0        0        6
#P-var   #M-var   #B-var   #C-var   #I-var
    0        0        0       16        0
```

Number of cases used for training: 46693
Number of split variables: 22
Number of cases excluded due to 0 weight or missing D: 0

Pruning by v-fold cross-validation, with v = 10
Selected tree is based on mean of CV estimates
Number of SE's for pruned tree: 0.5000

Simple node models
Estimated priors
Unit misclassification costs
Univariate split highest priority
Interaction and linear splits 2nd and 3rd priorities
Split values for N and S variables based on exhaustive search
Maximum number of split levels: 30
Minimum node sample size: 466
Top-ranked variables and chi-squared values at root node

```
 1  0.4620E+05   INTP
 2  0.1248E+05   WRK
 3  0.8637E+04   ANC
 4  0.9105E+03   RELSHIPP
 5  0.6668E+03   POVPIP
 6  0.6236E+03   NWAB
 7  0.5975E+03   MSP
 8  0.5837E+03   NWLK
 9  0.5683E+03   NWLA
10  0.5631E+03   OC
11  0.5611E+03   RC
12  0.3681E+03   MAR
13  0.3299E+03   DDRS
14  0.3271E+03   SCHL
15  0.2707E+03   AGEP
16  0.2562E+03   PERNP
17  0.2514E+03   DPHY
```

```
   18  0.2266E+03   WKL
   19  0.2230E+03   HINS3
   20  0.1964E+03   WKWN
   21  0.1928E+03   WKHP
   22  0.1726E+03   COW

Size and CV mean cost and SE of subtrees:
 Tree   #Tnodes  Mean Cost   SE(Mean)   BSE(Mean)  Median Cost  BSE(Median)
  0       3   4.283E-05   3.029E-05   2.722E-05   0.000E+00   2.398E-05
  1**     2   4.283E-05   3.029E-05   2.722E-05   0.000E+00   2.398E-05
  2       1   1.508E-01   1.656E-03   6.002E-05   1.508E-01   8.658E-05

0-SE tree based on mean is marked with * and has 2 terminal nodes
0-SE tree based on median is marked with + and has 2 terminal nodes
Selected-SE tree based on mean using naive SE is marked with **
Selected-SE tree based on mean using bootstrap SE is marked with --
Selected-SE tree based on median and bootstrap SE is marked with ++
* tree, ** tree, + tree, and ++ tree all the same

Following tree is based on mean CV with naive SE estimate (**).

Structure of final tree. Each terminal node is marked with a T.

Node cost is node misclassification cost divided by number of training cases
      Node    Total    Train    Predicted       Node    Split         Interacting
     label    cases    cases    class           cost    variables     variable
         1    46693    46693    D            1.508E-01 INTP
        2T     7037     7037    C            1.819E-05 -
        3T    39656    39656    D            5.366E-05 INTP

Number of terminal nodes of final tree: 2
Total number of nodes of final tree: 3
Second best split variable (based on curvature test) at root node is WRK

Classification tree:

Node 1: INTP = NA
  Node 2: C
Node 1: INTP /= NA
  Node 3: D

**********************************************************

Predictor means below are means of cases with no missing values.

Node 1: Intermediate node
```

A case goes into Node 2 if INTP = NA
INTP mean = 2555.4789

Class	Number	Posterior
C	7037	0.1507E+00
D	39654	0.8492E+00
T	2	0.4283E-04

Number of training cases misclassified = 7039
Predicted class is D

Node 2: Terminal node

Class	Number	Posterior
C	7037	0.1000E+01
D	0	0.1819E-04
T	0	0.9173E-09

Number of training cases misclassified = 0
Predicted class is C

Node 3: Terminal node

Class	Number	Posterior
C	0	0.3228E-05
D	39654	0.9999E+00
T	2	0.5043E-04

Number of training cases misclassified = 2
Predicted class is D

Classification matrix for training sample:

Predicted class	True class C	D	T
C	7037	0	0
D	0	39654	2
T	0	0	0
Total	7037	39654	2

Number of cases used for tree construction: 46693
Number misclassified: 2
Resubstitution estimate of mean misclassification cost: 0.42832973E-04

Observed and fitted values are stored in fit.txt
LaTeX code for tree is in code.tex
Elapsed time in seconds: 255.76

Output file of classification tree

Classification tree
Pruning by cross-validation
Data description file: descclass.txt
Training sample file: data.txt
Missing value code: NA
Records in data file start on line 2
5 N variables changed to S
D variable is FINTP
Number of records in data file: 46693
Length of longest entry in data file: 13
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Number of classes: 3
Training sample class proportions of D variable FINTP:

Class	#Cases	Proportion
C	7037	0.15070781
D	39654	0.84924935
T	2	0.00004283

Summary information for training sample of size 46693
d=dependent, b=split and fit cat variable using indicator variables,
c=split-only categorical, i=fit-only categorical (via indicators),
s=split-only numerical, n=split and fit numerical, f=fit-only numerical,
m=missing-value flag variable, p=periodic variable, w=weight

Column	Name		Minimum	Maximum	#Codes/ Levels/ Periods	#Missing
10	AGEP	s	15.00	95.00		
13	COW	c			9	11089
14	DDRS	c			2	
18	DPHY	c			2	
30	HINS3	c			2	
40	MAR	c			5	
57	NWAB	c			3	779
59	NWLA	c			3	779
60	NWLK	c			3	779
64	RELSHIPP	c			19	
68	SCHL	c			24	
74	WKHP	s	1.000	99.00		15014
75	WKL	c			3	779
76	WKWN	s	1.000	52.00		15014
77	WRK	c			2	4730
79	ANC	c			4	
97	MSP	c			6	

```
  101  OC        c                        2      2465
  104  PERNP     s   -8000.      0.8310E+06             779
  107  POVPIP    s    0.000       501.0                1752
  124  RC        c                        2      2465
  167  FINTP     d                        3

   Total  #cases w/   #missing
  #cases    miss. D  ord. vals   #X-var   #N-var   #F-var   #S-var
   46693          0      15859      266        0        0        5
  #P-var   #M-var   #B-var   #C-var   #I-var
       0        0        0       16        0
Number of cases used for training: 46693
Number of split variables: 21
Number of cases excluded due to 0 weight or missing D: 0

Pruning by v-fold cross-validation, with v = 10
Selected tree is based on mean of CV estimates
Number of SE's for pruned tree: 0.5000

Simple node models
Estimated priors
Unit misclassification costs
Univariate split highest priority
Interaction and linear splits 2nd and 3rd priorities
Split values for N and S variables based on exhaustive search
Maximum number of split levels: 30
Minimum node sample size: 466
Top-ranked variables and chi-squared values at root node
    1  0.1248E+05   WRK
    2  0.8637E+04   ANC
    3  0.9105E+03   RELSHIPP
    4  0.6668E+03   POVPIP
    5  0.6236E+03   NWAB
    6  0.5975E+03   MSP
    7  0.5837E+03   NWLK
    8  0.5683E+03   NWLA
    9  0.5631E+03   OC
   10  0.5611E+03   RC
   11  0.3681E+03   MAR
   12  0.3299E+03   DDRS
   13  0.3271E+03   SCHL
   14  0.2707E+03   AGEP
   15  0.2562E+03   PERNP
   16  0.2514E+03   DPHY
   17  0.2266E+03   WKL
   18  0.2230E+03   HINS3
```

19	0.1964E+03	WKWN
20	0.1928E+03	WKHP
21	0.1726E+03	COW

Size and CV mean cost and SE of subtrees:

Tree	#Tnodes	Mean Cost	SE(Mean)	BSE(Mean)	Median Cost	BSE(Median)
1	72	9.451E-02	1.354E-03	8.423E-04	9.467E-02	1.011E-03
2	70	9.451E-02	1.354E-03	8.423E-04	9.467E-02	1.011E-03
3	69	9.451E-02	1.354E-03	8.423E-04	9.467E-02	1.011E-03
4	68	9.451E-02	1.354E-03	8.423E-04	9.467E-02	1.011E-03
5	67	9.451E-02	1.354E-03	8.423E-04	9.467E-02	1.011E-03
6	66	9.451E-02	1.354E-03	8.423E-04	9.467E-02	1.011E-03
7	65	9.451E-02	1.354E-03	8.423E-04	9.467E-02	1.011E-03
8	62	9.451E-02	1.354E-03	8.423E-04	9.467E-02	1.011E-03
9	61	9.451E-02	1.354E-03	8.423E-04	9.467E-02	1.011E-03
10	60	9.451E-02	1.354E-03	8.423E-04	9.467E-02	1.011E-03
11	59	9.451E-02	1.354E-03	8.423E-04	9.467E-02	1.011E-03
12	58	9.451E-02	1.354E-03	8.423E-04	9.467E-02	1.011E-03
13	52	9.451E-02	1.354E-03	8.423E-04	9.467E-02	1.011E-03
14	50	9.451E-02	1.354E-03	8.423E-04	9.467E-02	1.011E-03
15	48	9.451E-02	1.354E-03	8.423E-04	9.467E-02	1.011E-03
16	47	9.451E-02	1.354E-03	8.423E-04	9.467E-02	1.011E-03
17	46	9.451E-02	1.354E-03	8.423E-04	9.467E-02	1.011E-03
18	45	9.451E-02	1.354E-03	8.423E-04	9.467E-02	1.011E-03
19	44	9.451E-02	1.354E-03	8.423E-04	9.467E-02	1.011E-03
20	41	9.451E-02	1.354E-03	8.423E-04	9.467E-02	1.011E-03
21	40	9.451E-02	1.354E-03	8.423E-04	9.467E-02	1.011E-03
22	39	9.451E-02	1.354E-03	8.423E-04	9.467E-02	1.011E-03
23	38	9.451E-02	1.354E-03	8.423E-04	9.467E-02	1.011E-03
24	37	9.451E-02	1.354E-03	8.423E-04	9.467E-02	1.011E-03
25	36	9.451E-02	1.354E-03	8.423E-04	9.467E-02	1.011E-03
26	35	9.451E-02	1.354E-03	8.423E-04	9.467E-02	1.011E-03
27	34	9.451E-02	1.354E-03	8.423E-04	9.467E-02	1.011E-03
28	33	9.451E-02	1.354E-03	8.423E-04	9.467E-02	1.011E-03
29	32	9.451E-02	1.354E-03	8.423E-04	9.467E-02	1.011E-03
30	31	9.451E-02	1.354E-03	8.423E-04	9.467E-02	1.011E-03
31	27	9.451E-02	1.354E-03	8.423E-04	9.467E-02	1.011E-03
32	25	9.451E-02	1.354E-03	8.423E-04	9.467E-02	1.011E-03
33	24	9.451E-02	1.354E-03	8.423E-04	9.467E-02	1.011E-03
34	23	9.451E-02	1.354E-03	8.423E-04	9.467E-02	1.011E-03
35	22	9.451E-02	1.354E-03	8.423E-04	9.467E-02	1.011E-03
36	21	9.451E-02	1.354E-03	8.423E-04	9.467E-02	1.011E-03
37	6	9.451E-02	1.354E-03	8.423E-04	9.467E-02	1.011E-03
38	5	9.451E-02	1.354E-03	8.423E-04	9.467E-02	1.011E-03
39++	4	9.451E-02	1.354E-03	8.423E-04	9.467E-02	1.011E-03
40**	3	9.490E-02	1.356E-03	1.017E-03	9.520E-02	1.311E-03

41	2	1.088E-01	1.441E-03	9.203E-04	1.094E-01	9.910E-04
42	1	1.508E-01	1.656E-03	1.159E-04	1.506E-01	1.614E-04

0-SE tree based on mean is marked with * and has 4 terminal nodes
0-SE tree based on median is marked with + and has 4 terminal nodes
Selected-SE tree based on mean using naive SE is marked with **
Selected-SE tree based on mean using bootstrap SE is marked with --
Selected-SE tree based on median and bootstrap SE is marked with ++
* tree same as + tree
** tree same as -- tree
+ tree same as ++ tree
* tree same as ++ tree

Following tree is based on mean CV with naive SE estimate (**).

Structure of final tree. Each terminal node is marked with a T.

Node cost is node misclassification cost divided by number of training cases

Node label	Total cases	Train cases	Predicted class	Node cost	Split variables	Interacting variable
1	46693	46693	D	1.508E-01	WRK	
2	4730	4730	C	2.930E-01	NWLK	
4T	3758	3758	C	1.525E-01	RELSHIPP	
5T	972	972	D	1.636E-01	POVPIP	
3T	41963	41963	D	8.805E-02	ANC	

Number of terminal nodes of final tree: 3
Total number of nodes of final tree: 5
Second best split variable (based on curvature test) at root node is ANC

Classification tree:
For categorical variable splits, values not in training data go to the right

Node 1: WRK = "NA"
  Node 2: NWLK = "3"
    Node 4: C
  Node 2: NWLK /= "3"
    Node 5: D
Node 1: WRK /= "NA"
  Node 3: D

****

Predictor means below are means of cases with no missing values.

Node 1: Intermediate node

A case goes into Node 2 if WRK = "NA"
WRK mode = "1"
Class Number Posterior
C 7037 0.1507E+00
D 39654 0.8492E+00
T 2 0.4283E-04
Number of training cases misclassified = 7039
Predicted class is D

Node 2: Intermediate node
A case goes into Node 4 if NWLK = "3"
NWLK mode = "3"
Class Number Posterior
C 3344 0.7070E+00
D 1386 0.2930E+00
T 0 0.9173E-09
Number of training cases misclassified = 1386
Predicted class is C

Node 4: Terminal node
Class Number Posterior
C 3185 0.8475E+00
D 573 0.1525E+00
T 0 0.9173E-09
Number of training cases misclassified = 573
Predicted class is C

Node 5: Terminal node
Class Number Posterior
C 159 0.1636E+00
D 813 0.8364E+00
T 0 0.9173E-09
Number of training cases misclassified = 159
Predicted class is D

Node 3: Terminal node
Class Number Posterior
C 3693 0.8801E-01
D 38268 0.9119E+00
T 2 0.4766E-04
Number of training cases misclassified = 3695
Predicted class is D

Classification matrix for training sample:
Predicted True class

class	C	D	T
C	3185	573	0
D	3852	39081	2
T	0	0	0
Total	7037	39654	2

Number of cases used for tree construction: 46693
Number misclassified: 4427
Resubstitution estimate of mean misclassification cost: 0.94810785E-01

Observed and fitted values are stored in fit.txt
LaTeX code for tree is in code.tex
Elapsed time in seconds: 295.42

Output file of classification tree at the second best split

Classification tree
Pruning by cross-validation
Data description file: descclass.txt
Training sample file: data.txt
Missing value code: NA
Records in data file start on line 2
5 N variables changed to S
D variable is FINTP
Number of records in data file: 46693
Length of longest entry in data file: 13
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Number of classes: 3
Training sample class proportions of D variable FINTP:

Class	#Cases	Proportion
C	7037	0.15070781
D	39654	0.84924935
T	2	0.00004283

Summary information for training sample of size 46693
d=dependent, b=split and fit cat variable using indicator variables,
c=split-only categorical, i=fit-only categorical (via indicators),
s=split-only numerical, n=split and fit numerical, f=fit-only numerical,
m=missing-value flag variable, p=periodic variable, w=weight

Column	Name		Minimum	Maximum	#Codes/ Levels/ Periods	#Missing
10	AGEP	s	15.00	95.00		
13	COW	c			9	11089
14	DDRS	c			2	
18	DPHY	c			2	
30	HINS3	c			2	
40	MAR	c			5	
57	NWAB	c			3	779
59	NWLA	c			3	779
60	NWLK	c			3	779
64	RELSHIPP	c			19	
68	SCHL	c			24	
74	WKHP	s	1.000	99.00		15014
75	WKL	c			3	779
76	WKWN	s	1.000	52.00		15014
77	WRK	c			2	4730
79	ANC	c			4	
97	MSP	c			6	

```
  101  OC        c                        2      2465
  104  PERNP     s   -8000.      0.8310E+06             779
  107  POVPIP    s    0.000       501.0                1752
  124  RC        c                        2      2465
  167  FINTP     d                         3

   Total  #cases w/   #missing
  #cases    miss. D  ord. vals   #X-var   #N-var   #F-var   #S-var
   46693          0      15859      266        0        0        5
  #P-var   #M-var   #B-var   #C-var   #I-var
       0        0        0       16        0
Number of cases used for training: 46693
Number of split variables: 21
Number of cases excluded due to 0 weight or missing D: 0

Pruning by v-fold cross-validation, with v = 10
Selected tree is based on mean of CV estimates
Number of SE's for pruned tree: 0.5000

Simple node models
Estimated priors
Unit misclassification costs
Univariate split highest priority
Interaction and linear splits 2nd and 3rd priorities
Split values for N and S variables based on exhaustive search
Maximum number of split levels: 30
Minimum node sample size: 466
Rank of top variable to split root node is 2
Top-ranked variables and chi-squared values at root node
    1  0.1248E+05   WRK
    2  0.8637E+04   ANC
    3  0.9105E+03   RELSHIPP
    4  0.6668E+03   POVPIP
    5  0.6236E+03   NWAB
    6  0.5975E+03   MSP
    7  0.5837E+03   NWLK
    8  0.5683E+03   NWLA
    9  0.5631E+03   OC
   10  0.5611E+03   RC
   11  0.3681E+03   MAR
   12  0.3299E+03   DDRS
   13  0.3271E+03   SCHL
   14  0.2707E+03   AGEP
   15  0.2562E+03   PERNP
   16  0.2514E+03   DPHY
   17  0.2266E+03   WKL
```

18	0.2230E+03	HINS3
19	0.1964E+03	WKWN
20	0.1928E+03	WKHP
21	0.1726E+03	COW

Size and CV mean cost and SE of subtrees:

Tree	#Tnodes	Mean Cost	SE(Mean)	BSE(Mean)	Median Cost	BSE(Median)
1	73	9.607E-02	1.364E-03	1.148E-03	9.681E-02	1.584E-03
2	72	9.607E-02	1.364E-03	1.148E-03	9.681E-02	1.584E-03
3	71	9.607E-02	1.364E-03	1.148E-03	9.681E-02	1.584E-03
4	70	9.607E-02	1.364E-03	1.148E-03	9.681E-02	1.584E-03
5	69	9.607E-02	1.364E-03	1.148E-03	9.681E-02	1.584E-03
6	68	9.607E-02	1.364E-03	1.148E-03	9.681E-02	1.584E-03
7	65	9.607E-02	1.364E-03	1.148E-03	9.681E-02	1.584E-03
8	64	9.607E-02	1.364E-03	1.148E-03	9.681E-02	1.584E-03
9	63	9.607E-02	1.364E-03	1.148E-03	9.681E-02	1.584E-03
10	62	9.607E-02	1.364E-03	1.148E-03	9.681E-02	1.584E-03
11	61	9.607E-02	1.364E-03	1.148E-03	9.681E-02	1.584E-03
12	55	9.607E-02	1.364E-03	1.148E-03	9.681E-02	1.584E-03
13	53	9.607E-02	1.364E-03	1.148E-03	9.681E-02	1.584E-03
14	51	9.607E-02	1.364E-03	1.148E-03	9.681E-02	1.584E-03
15	50	9.607E-02	1.364E-03	1.148E-03	9.681E-02	1.584E-03
16	49	9.607E-02	1.364E-03	1.148E-03	9.681E-02	1.584E-03
17	48	9.607E-02	1.364E-03	1.148E-03	9.681E-02	1.584E-03
18	47	9.607E-02	1.364E-03	1.148E-03	9.681E-02	1.584E-03
19	44	9.607E-02	1.364E-03	1.148E-03	9.681E-02	1.584E-03
20	43	9.607E-02	1.364E-03	1.148E-03	9.681E-02	1.584E-03
21	42	9.607E-02	1.364E-03	1.148E-03	9.681E-02	1.584E-03
22	41	9.607E-02	1.364E-03	1.148E-03	9.681E-02	1.584E-03
23	40	9.607E-02	1.364E-03	1.148E-03	9.681E-02	1.584E-03
24	39	9.607E-02	1.364E-03	1.148E-03	9.681E-02	1.584E-03
25	38	9.607E-02	1.364E-03	1.148E-03	9.681E-02	1.584E-03
26	37	9.607E-02	1.364E-03	1.148E-03	9.681E-02	1.584E-03
27	36	9.607E-02	1.364E-03	1.148E-03	9.681E-02	1.584E-03
28	35	9.607E-02	1.364E-03	1.148E-03	9.681E-02	1.584E-03
29	34	9.607E-02	1.364E-03	1.148E-03	9.681E-02	1.584E-03
30	30	9.607E-02	1.364E-03	1.148E-03	9.681E-02	1.584E-03
31	28	9.607E-02	1.364E-03	1.148E-03	9.681E-02	1.584E-03
32	27	9.607E-02	1.364E-03	1.148E-03	9.681E-02	1.584E-03
33	26	9.607E-02	1.364E-03	1.148E-03	9.681E-02	1.584E-03
34	25	9.607E-02	1.364E-03	1.148E-03	9.681E-02	1.584E-03
35	24	9.607E-02	1.364E-03	1.148E-03	9.681E-02	1.584E-03
36	9	9.607E-02	1.364E-03	1.148E-03	9.681E-02	1.584E-03
37	8	9.607E-02	1.364E-03	1.148E-03	9.681E-02	1.584E-03
38**	6	9.607E-02	1.364E-03	1.148E-03	9.681E-02	1.584E-03
39	3	9.860E-02	1.380E-03	6.257E-04	9.884E-02	7.073E-04

40 1 1.508E-01 1.656E-03 1.159E-04 1.506E-01 1.614E-04

0-SE tree based on mean is marked with * and has 6 terminal nodes
0-SE tree based on median is marked with + and has 6 terminal nodes
Selected-SE tree based on mean using naive SE is marked with **
Selected-SE tree based on mean using bootstrap SE is marked with --
Selected-SE tree based on median and bootstrap SE is marked with ++
* tree, ** tree, + tree, and ++ tree all the same

Following tree is based on mean CV with naive SE estimate (**).

Structure of final tree. Each terminal node is marked with a T.

Node cost is node misclassification cost divided by number of training cases

Node label	Total cases	Train cases	Predicted class	Node cost	Split variables	Interacting variable
1	46693	46693	D	1.508E-01	ANC	
2	7123	7123	C	4.824E-01	WRK	
4T	3041	3041	C	9.965E-02	WKL	
5T	4082	4082	D	2.325E-01	AGEP	
3	39570	39570	D	8.471E-02	WRK	
6	1689	1689	D	3.588E-01	RC	
12T	605	605	D	4.794E-02	-	
13	1084	1084	C	4.677E-01	RC	
26T	580	580	C	3.310E-01	-	
27T	504	504	D	3.750E-01	-	
7T	37881	37881	D	7.249E-02	AGEP	

Number of terminal nodes of final tree: 6
Total number of nodes of final tree: 11

Classification tree:
For categorical variable splits, values not in training data go to the right

Node 1: ANC = "4"
  Node 2: WRK = "NA"
    Node 4: C
  Node 2: WRK /= "NA"
    Node 5: D
Node 1: ANC /= "4"
  Node 3: WRK = "NA"
    Node 6: RC = "1"
      Node 12: D
    Node 6: RC /= "1"
      Node 13: RC = "0"
        Node 26: C

```
      Node 13: RC /= "0"
        Node 27: D
    Node 3: WRK /= "NA"
      Node 7: D

 ************************************************************

 Predictor means below are means of cases with no missing values.

Node 1: Intermediate node
 A case goes into Node 2 if ANC = "4"
 ANC mode = "1"
 Class      Number   Posterior
 C           7037  0.1507E+00
 D          39654  0.8492E+00
 T              2  0.4283E-04
 Number of training cases misclassified = 7039
 Predicted class is D
 ---------------------------
Node 2: Intermediate node
 A case goes into Node 4 if WRK = "NA"
 WRK mode = "NA"
 Class      Number   Posterior
 C           3687  0.5176E+00
 D           3436  0.4824E+00
 T              0  0.9173E-09
 Number of training cases misclassified = 3436
 Predicted class is C
 ---------------------------
Node 4: Terminal node
 Class      Number   Posterior
 C           2738  0.9003E+00
 D            303  0.9965E-01
 T              0  0.9173E-09
 Number of training cases misclassified = 303
 Predicted class is C
 ---------------------------
Node 5: Terminal node
 Class      Number   Posterior
 C            949  0.2325E+00
 D           3133  0.7675E+00
 T              0  0.9173E-09
 Number of training cases misclassified = 949
 Predicted class is D
 ---------------------------
Node 3: Intermediate node
```

A case goes into Node 6 if WRK = "NA"
WRK mode = "1"
Class Number Posterior
C 3350 0.8466E-01
D 36218 0.9153E+00
T 2 0.5054E-04
Number of training cases misclassified = 3352
Predicted class is D

Node 6: Intermediate node
A case goes into Node 12 if RC = "1"
RC mode = "1"
Class Number Posterior
C 606 0.3588E+00
D 1083 0.6412E+00
T 0 0.9173E-09
Number of training cases misclassified = 606
Predicted class is D

Node 12: Terminal node
Class Number Posterior
C 29 0.4794E-01
D 576 0.9521E+00
T 0 0.9173E-09
Number of training cases misclassified = 29
Predicted class is D

Node 13: Intermediate node
A case goes into Node 26 if RC = "0"
RC mode = "0"
Class Number Posterior
C 577 0.5323E+00
D 507 0.4677E+00
T 0 0.9173E-09
Number of training cases misclassified = 507
Predicted class is C

Node 26: Terminal node
Class Number Posterior
C 388 0.6690E+00
D 192 0.3310E+00
T 0 0.9173E-09
Number of training cases misclassified = 192
Predicted class is C

Node 27: Terminal node

```
 Class      Number   Posterior
 C           189  0.3750E+00
 D           315  0.6250E+00
 T             0  0.9173E-09
 Number of training cases misclassified = 189
 Predicted class is D
 ---------------------------
Node 7: Terminal node
 Class      Number   Posterior
 C          2744  0.7244E-01
 D         35135  0.9275E+00
 T             2  0.5280E-04
 Number of training cases misclassified = 2746
 Predicted class is D
 ---------------------------

 Classification matrix for training sample:
 Predicted      True class
 class            C        D        T
 C             3126      495        0
 D             3911    39159        2
 T                0        0        0
 Total         7037    39654        2

 Number of cases used for tree construction: 46693
 Number misclassified: 4408
 Resubstitution estimate of mean misclassification cost: 0.94403872E-01

 Observed and fitted values are stored in fit.txt
 LaTeX code for tree is in code.tex
 Elapsed time in seconds: 291.31
```

Output file of regression piecewise constant tree

Least squares regression tree
Pruning by cross-validation
Data description file: descreg.txt
Training sample file: data.txt
Missing value code: NA
Records in data file start on line 2
13 N variables changed to S
D variable is INTP
Piecewise constant model
Number of records in data file: 46693
Length of longest entry in data file: 13
Missing values found in D variable
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Smallest and largest positive weights are 1.0000E+00 and 2.4080E+03

Summary information for training sample of size 39656 (excluding observations with non-positive weight or missing values in d, e, t, r or z variables)
d=dependent, b=split and fit cat variable using indicator variables,
c=split-only categorical, i=fit-only categorical (via indicators),
s=split-only numerical, n=split and fit numerical, f=fit-only numerical,
m=missing-value flag variable, p=periodic variable, w=weight

Column	Name		Minimum	Maximum	#Codes/ Levels/ Periods	#Missing
9	PWGTP	w	1.000	2408.		
10	AGEP	s	15.00	95.00		
13	COW	c			9	8906
23	FER	c			2	30173
24	GCL	c			2	7529
27	HIMRKS	c			3	
28	HINS1	c			2	
29	HINS2	c			2	
30	HINS3	c			2	
35	INTP	d	-1300.	0.2540E+06		
36	JWMNP	s	1.000	163.0		17020
38	JWTRNS	c			12	15212
40	MAR	c			5	
41	MARHD	c			2	10258
42	MARHM	c			2	10258
43	MARHT	c			3	10258
44	MARHW	c			2	10258
45	MARHYP	s	1940.	2019.		10258
57	NWAB	c			3	672

```
 59  NWLA       c                          3       672
 60  NWLK       c                          3       672
 64  RELSHIPP   c                         19
 65  RETP       s    0.000       0.1420E+06
 66  SCH        c                          3
 68  SCHL       c                         24
 70  SEX        c                          2
 72  SSP        s    0.000       0.3690E+05
 73  WAGP       s    0.000       0.4760E+06
 74  WKHP       s    1.000       99.00              12232
 75  WKL        c                          3       672
 76  WKWN       s    1.000       52.00              12232
 77  WRK        c                          2      1386
 86  ESR        c                          5       672
 92  JWAP       s    1.000       285.0              17020
 93  JWDP       s    1.000       150.0              17020
 97  MSP        c                          6
101  OC         c                          2      1665
104  PERNP      s   -8000.       0.8310E+06             672
105  PINCP      s   -9300.       0.8590E+06
107  POVPIP     s    0.000       501.0              1126
111  PUBCOV     c                          2
119  RACNH      c                          2
124  RC         c                          2      1665
167  FINTP      c                          3

 Total  #cases w/   #missing
#cases    miss. D  ord. vals   #X-var   #N-var   #F-var   #S-var
 46693       7037      30071      244        0        0       13
#P-var   #M-var   #B-var   #C-var   #I-var
    0        0        0       29        0
Weight variable PWGTP in column: 9
Number of cases used for training: 39656
Number of split variables: 42
Number of cases excluded due to 0 weight or missing D: 7037

Pruning by v-fold cross-validation, with v = 10
Selected tree is based on mean of CV estimates
Number of SE's for pruned tree: 0.5000

Weighted error estimates used for pruning
Nodewise interaction tests on all variables
Split values for N and S variables based on exhaustive search
Maximum number of split levels: 30
Minimum node sample size: 396
Top-ranked variables and chi-squared values at root node
```

1	0.2023E+04	AGEP
2	0.1564E+04	PINCP
3	0.1433E+04	HINS3
4	0.1413E+04	MARHYP
5	0.1254E+04	SSP
6	0.1069E+04	RETP
7	0.1009E+04	HIMRKS
8	0.1003E+04	HINS2
9	0.7966E+03	MAR
10	0.7885E+03	MSP
11	0.7858E+03	RELSHIPP
12	0.7561E+03	WAGP
13	0.7018E+03	COW
14	0.6975E+03	WKL
15	0.6659E+03	PERNP
16	0.6544E+03	GCL
17	0.6512E+03	ESR
18	0.6402E+03	PUBCOV
19	0.6384E+03	FER
20	0.6332E+03	WRK
21	0.6227E+03	NWLK
22	0.6217E+03	WKHP
23	0.5839E+03	POVPIP
24	0.5668E+03	WKWN
25	0.5633E+03	MARHM
26	0.5356E+03	MARHW
27	0.5337E+03	MARHD
28	0.5151E+03	MARHT
29	0.4983E+03	SCHL
30	0.4802E+03	NWLA
31	0.4802E+03	JWDP
32	0.4758E+03	JWAP
33	0.4705E+03	JWMNP
34	0.4638E+03	NWAB
35	0.4267E+03	JWTRNS
36	0.3264E+03	SCH
37	0.3082E+03	HINS1
38	0.3042E+03	RC
39	0.2953E+03	OC
40	0.1650E+03	SEX

Size and CV MSE and SE of subtrees:

Tree	#Tnodes	Mean MSE	SE(Mean)	BSE(Mean)	Median MSE	BSE(Median)
1	78	2.130E+10	1.625E+09	7.569E+08	2.118E+10	1.304E+09
2	76	2.130E+10	1.625E+09	7.569E+08	2.118E+10	1.304E+09
3	75	2.130E+10	1.625E+09	7.569E+08	2.118E+10	1.304E+09

4	74	2.130E+10	1.625E+09	7.569E+08	2.118E+10	1.304E+09
5	73	2.130E+10	1.625E+09	7.569E+08	2.118E+10	1.304E+09
6	72	2.130E+10	1.625E+09	7.569E+08	2.118E+10	1.304E+09
7	70	2.130E+10	1.625E+09	7.569E+08	2.118E+10	1.304E+09
8	68	2.130E+10	1.625E+09	7.569E+08	2.118E+10	1.304E+09
9	66	2.130E+10	1.625E+09	7.569E+08	2.118E+10	1.304E+09
10	64	2.130E+10	1.625E+09	7.569E+08	2.118E+10	1.304E+09
11	62	2.130E+10	1.625E+09	7.569E+08	2.118E+10	1.304E+09
12	61	2.130E+10	1.625E+09	7.569E+08	2.118E+10	1.304E+09
13	60	2.130E+10	1.625E+09	7.569E+08	2.118E+10	1.304E+09
14	59	2.130E+10	1.625E+09	7.569E+08	2.118E+10	1.304E+09
15	58	2.130E+10	1.625E+09	7.569E+08	2.118E+10	1.304E+09
16	57	2.130E+10	1.625E+09	7.569E+08	2.118E+10	1.304E+09
17	56	2.130E+10	1.625E+09	7.569E+08	2.118E+10	1.304E+09
18	55	2.130E+10	1.625E+09	7.569E+08	2.118E+10	1.304E+09
19	54	2.130E+10	1.625E+09	7.569E+08	2.118E+10	1.304E+09
20	51	2.130E+10	1.625E+09	7.569E+08	2.118E+10	1.304E+09
21	50	2.130E+10	1.625E+09	7.569E+08	2.118E+10	1.304E+09
22	48	2.130E+10	1.625E+09	7.569E+08	2.118E+10	1.304E+09
23	47	2.130E+10	1.625E+09	7.568E+08	2.118E+10	1.304E+09
24*	44	2.130E+10	1.625E+09	7.568E+08	2.118E+10	1.304E+09
25+	43	2.130E+10	1.625E+09	7.568E+08	2.118E+10	1.304E+09
26	41	2.130E+10	1.625E+09	7.568E+08	2.118E+10	1.304E+09
27	39	2.130E+10	1.625E+09	7.569E+08	2.118E+10	1.304E+09
28	36	2.130E+10	1.625E+09	7.569E+08	2.118E+10	1.304E+09
29	33	2.130E+10	1.625E+09	7.570E+08	2.118E+10	1.304E+09
30	32	2.130E+10	1.625E+09	7.570E+08	2.118E+10	1.304E+09
31	31	2.130E+10	1.625E+09	7.570E+08	2.118E+10	1.304E+09
32	30	2.130E+10	1.625E+09	7.572E+08	2.118E+10	1.305E+09
33	28	2.130E+10	1.625E+09	7.572E+08	2.118E+10	1.305E+09
34	26	2.130E+10	1.625E+09	7.569E+08	2.118E+10	1.304E+09
35	25	2.130E+10	1.625E+09	7.569E+08	2.118E+10	1.304E+09
36	24	2.130E+10	1.625E+09	7.569E+08	2.118E+10	1.304E+09
37	22	2.131E+10	1.625E+09	7.570E+08	2.118E+10	1.303E+09
38	21	2.131E+10	1.625E+09	7.562E+08	2.119E+10	1.302E+09
39	20	2.131E+10	1.625E+09	7.561E+08	2.119E+10	1.302E+09
40	19	2.131E+10	1.625E+09	7.565E+08	2.119E+10	1.303E+09
41	17	2.131E+10	1.625E+09	7.550E+08	2.119E+10	1.303E+09
42	16	2.132E+10	1.625E+09	7.540E+08	2.119E+10	1.307E+09
43	12	2.133E+10	1.625E+09	7.537E+08	2.120E+10	1.306E+09
44	11	2.133E+10	1.625E+09	7.543E+08	2.120E+10	1.310E+09
45	10	2.133E+10	1.625E+09	7.553E+08	2.120E+10	1.311E+09
46	8	2.136E+10	1.625E+09	7.525E+08	2.127E+10	1.292E+09
47--	7	2.148E+10	1.626E+09	7.716E+08	2.138E+10	1.367E+09
48**	4	2.177E+10	1.627E+09	8.036E+08	2.171E+10	1.394E+09
49	3	2.299E+10	1.901E+09	9.842E+08	2.306E+10	1.564E+09

```
  50       1   2.708E+10   2.422E+09   1.300E+09   2.839E+10   1.542E+09
```

0-SE tree based on mean is marked with * and has 44 terminal nodes
0-SE tree based on median is marked with + and has 43 terminal nodes
Selected-SE tree based on mean using naive SE is marked with **
Selected-SE tree based on mean using bootstrap SE is marked with --
Selected-SE tree based on median and bootstrap SE is marked with ++
** tree same as ++ tree

Following tree is based on mean CV with naive SE estimate (**).

Structure of final tree. Each terminal node is marked with a T.

D-mean is weighted mean of INTP in the node
Cases fit give the number of cases used to fit node
MSE is residual sum of squares divided by number of cases in node

Node label	Total cases	Cases fit	Matrix rank	Node D-mean	Node MSE	Split variable	Interacting variable
1	39656	39656	1	2.008E+03	2.708E+10	AGEP	
2	27376	27376	1	1.014E+03	1.510E+10	PINCP	
4T	26975	26975	1	4.755E+02	1.515E+09	PINCP	
5T	401	401	1	3.438E+04	7.983E+11	-	
3	12280	12280	1	5.125E+03	5.280E+10	PINCP	
6T	11884	11884	1	2.444E+03	5.473E+09	PINCP	
7T	396	396	1	7.258E+04	1.044E+12	-	

Number of terminal nodes of final tree: 4
Total number of nodes of final tree: 7
Second best split variable (based on curvature test) at root node is PINCP

Regression tree:

```
Node 1: AGEP <= 61.500000
  Node 2: PINCP <= 250625.00
    Node 4: INTP-mean = 475.50051
  Node 2: PINCP > 250625.00 or NA
    Node 5: INTP-mean = 34379.013
Node 1: AGEP > 61.500000 or NA
  Node 3: PINCP <= 155750.00
    Node 6: INTP-mean = 2444.4788
  Node 3: PINCP > 155750.00 or NA
    Node 7: INTP-mean = 72580.058
```

****

Predictor means below are weighted means of cases with no missing values.

WARNING: p-values below not adjusted for split search. For a bootstrap solution see:

1. Loh et al. (2016), "Identification of subgroups with differential treatment effects for longitudinal and multiresponse variables", Statistics in Medicine, v.35, 4837-4855.

2. Loh et al. (2019), "Subgroups from regression trees with adjustment for prognostic effects and post-selection inference", Statistics in Medicine, v.38, 545-557.

3. Loh and Zhou (2020), "The GUIDE approach to subgroup identification", in "Design and Analysis of Subgroups with Biopharmaceutical Applications", Springer, pp.147-165.

Node 1: Intermediate node
A case goes into Node 2 if AGEP <= 61.500000
AGEP mean = 46.273292
Coefficients of least squares regression function:

Regressor	Coefficient	t-stat	p-value
Constant	2008.	15.87	0.000

INTP mean = 2007.62

Node 2: Intermediate node
A case goes into Node 4 if PINCP <= 250625.00
PINCP mean = 50525.232

Node 4: Terminal node
Coefficients of least squares regression functions:

Regressor	Coefficient	t-stat	p-value
Constant	475.5	13.77	0.7994E-14

INTP mean = 475.501

Node 5: Terminal node
Coefficients of least squares regression functions:

Regressor	Coefficient	t-stat	p-value
Constant	0.3438E+05	5.977	0.5021E-08

INTP mean = 34379.0

Node 3: Intermediate node
A case goes into Node 6 if PINCP <= 155750.00
PINCP mean = 45569.851

Node 6: Terminal node
Coefficients of least squares regression functions:

Regressor	Coefficient	t-stat	p-value
Constant	2444.	21.51	0.000

INTP mean = 2444.48

Node 7: Terminal node
Coefficients of least squares regression functions:
Regressor Coefficient t-stat p-value
Constant 0.7258E+05 9.790 0.000
INTP mean = 72580.1

Proportion of variance (R-squared) explained by tree model: 0.2197

Observed and fitted values are stored in fit.txt
LaTeX code for tree is in code.tex
Elapsed time in seconds: 320.03

Output file of regression piecewise constant tree at the second best split

Least squares regression tree
Pruning by cross-validation
Data description file: descreg.txt
Training sample file: data.txt
Missing value code: NA
Records in data file start on line 2
13 N variables changed to S
D variable is INTP
Piecewise constant model
Number of records in data file: 46693
Length of longest entry in data file: 13
Missing values found in D variable
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Smallest and largest positive weights are 1.0000E+00 and 2.4080E+03

Summary information for training sample of size 39656 (excluding observations with non-positive weight or missing values in d, e, t, r or z variables)
d=dependent, b=split and fit cat variable using indicator variables,
c=split-only categorical, i=fit-only categorical (via indicators),
s=split-only numerical, n=split and fit numerical, f=fit-only numerical,
m=missing-value flag variable, p=periodic variable, w=weight

Column	Name		Minimum	Maximum	#Codes/ Levels/ Periods	#Missing
9	PWGTP	w	1.000	2408.		
10	AGEP	s	15.00	95.00		
13	COW	c			9	8906
23	FER	c			2	30173
24	GCL	c			2	7529
27	HIMRKS	c			3	
28	HINS1	c			2	
29	HINS2	c			2	
30	HINS3	c			2	
35	INTP	d	-1300.	0.2540E+06		
36	JWMNP	s	1.000	163.0		17020
38	JWTRNS	c			12	15212
40	MAR	c			5	
41	MARHD	c			2	10258
42	MARHM	c			2	10258
43	MARHT	c			3	10258
44	MARHW	c			2	10258
45	MARHYP	s	1940.	2019.		10258
57	NWAB	c			3	672

```
 59  NWLA       c                          3       672
 60  NWLK       c                          3       672
 64  RELSHIPP   c                         19
 65  RETP       s    0.000       0.1420E+06
 66  SCH        c                          3
 68  SCHL       c                         24
 70  SEX        c                          2
 72  SSP        s    0.000       0.3690E+05
 73  WAGP       s    0.000       0.4760E+06
 74  WKHP       s    1.000       99.00              12232
 75  WKL        c                          3       672
 76  WKWN       s    1.000       52.00              12232
 77  WRK        c                          2      1386
 86  ESR        c                          5       672
 92  JWAP       s    1.000       285.0              17020
 93  JWDP       s    1.000       150.0              17020
 97  MSP        c                          6
101  OC         c                          2      1665
104  PERNP      s   -8000.       0.8310E+06          672
105  PINCP      s   -9300.       0.8590E+06
107  POVPIP     s    0.000       501.0              1126
111  PUBCOV     c                          2
119  RACNH      c                          2
124  RC         c                          2      1665
167  FINTP      c                          3

 Total  #cases w/   #missing
#cases    miss. D  ord. vals   #X-var   #N-var   #F-var   #S-var
 46693       7037      30071      244        0        0       13
#P-var   #M-var   #B-var   #C-var   #I-var
    0        0        0       29        0
Weight variable PWGTP in column: 9
Number of cases used for training: 39656
Number of split variables: 42
Number of cases excluded due to 0 weight or missing D: 7037

Pruning by v-fold cross-validation, with v = 10
Selected tree is based on mean of CV estimates
Number of SE's for pruned tree: 0.5000

Weighted error estimates used for pruning
Nodewise interaction tests on all variables
Split values for N and S variables based on exhaustive search
Maximum number of split levels: 30
Minimum node sample size: 396
Rank of top variable to split root node is 2
```

Top-ranked variables and chi-squared values at root node

1	0.2023E+04	AGEP
2	0.1564E+04	PINCP
3	0.1433E+04	HINS3
4	0.1413E+04	MARHYP
5	0.1254E+04	SSP
6	0.1069E+04	RETP
7	0.1009E+04	HIMRKS
8	0.1003E+04	HINS2
9	0.7966E+03	MAR
10	0.7885E+03	MSP
11	0.7858E+03	RELSHIPP
12	0.7561E+03	WAGP
13	0.7018E+03	COW
14	0.6975E+03	WKL
15	0.6659E+03	PERNP
16	0.6544E+03	GCL
17	0.6512E+03	ESR
18	0.6402E+03	PUBCOV
19	0.6384E+03	FER
20	0.6332E+03	WRK
21	0.6227E+03	NWLK
22	0.6217E+03	WKHP
23	0.5839E+03	POVPIP
24	0.5668E+03	WKWN
25	0.5633E+03	MARHM
26	0.5356E+03	MARHW
27	0.5337E+03	MARHD
28	0.5151E+03	MARHT
29	0.4983E+03	SCHL
30	0.4802E+03	NWLA
31	0.4802E+03	JWDP
32	0.4758E+03	JWAP
33	0.4705E+03	JWMNP
34	0.4638E+03	NWAB
35	0.4267E+03	JWTRNS
36	0.3264E+03	SCH
37	0.3082E+03	HINS1
38	0.3042E+03	RC
39	0.2953E+03	OC
40	0.1650E+03	SEX

Size and CV MSE and SE of subtrees:

Tree	#Tnodes	Mean MSE	SE(Mean)	BSE(Mean)	Median MSE	BSE(Median)
1	76	2.030E+10	1.403E+09	8.829E+08	2.097E+10	1.052E+09
2	75	2.030E+10	1.403E+09	8.829E+08	2.097E+10	1.052E+09

3	74	2.030E+10	1.403E+09	8.829E+08	2.097E+10	1.052E+09
4	73	2.030E+10	1.403E+09	8.829E+08	2.097E+10	1.052E+09
5	71	2.030E+10	1.403E+09	8.829E+08	2.097E+10	1.052E+09
6*	70	2.030E+10	1.403E+09	8.829E+08	2.097E+10	1.052E+09
7	69	2.030E+10	1.403E+09	8.829E+08	2.097E+10	1.052E+09
8	68	2.030E+10	1.403E+09	8.829E+08	2.097E+10	1.052E+09
9	66	2.030E+10	1.403E+09	8.829E+08	2.097E+10	1.052E+09
10	63	2.030E+10	1.403E+09	8.829E+08	2.097E+10	1.052E+09
11	61	2.030E+10	1.403E+09	8.829E+08	2.097E+10	1.052E+09
12	60	2.030E+10	1.403E+09	8.829E+08	2.097E+10	1.052E+09
13+	58	2.030E+10	1.403E+09	8.829E+08	2.097E+10	1.052E+09
14	57	2.030E+10	1.403E+09	8.829E+08	2.097E+10	1.052E+09
15	56	2.030E+10	1.403E+09	8.829E+08	2.097E+10	1.052E+09
16	55	2.030E+10	1.403E+09	8.829E+08	2.097E+10	1.052E+09
17	53	2.030E+10	1.403E+09	8.829E+08	2.097E+10	1.052E+09
18	51	2.030E+10	1.403E+09	8.829E+08	2.097E+10	1.052E+09
19	49	2.030E+10	1.403E+09	8.829E+08	2.097E+10	1.052E+09
20	48	2.030E+10	1.403E+09	8.829E+08	2.097E+10	1.052E+09
21	44	2.030E+10	1.403E+09	8.829E+08	2.097E+10	1.052E+09
22	42	2.030E+10	1.403E+09	8.829E+08	2.097E+10	1.052E+09
23	41	2.030E+10	1.403E+09	8.829E+08	2.097E+10	1.052E+09
24	40	2.030E+10	1.403E+09	8.829E+08	2.097E+10	1.052E+09
25	38	2.030E+10	1.403E+09	8.829E+08	2.097E+10	1.052E+09
26	37	2.030E+10	1.403E+09	8.830E+08	2.097E+10	1.052E+09
27	35	2.030E+10	1.403E+09	8.830E+08	2.097E+10	1.052E+09
28	33	2.030E+10	1.403E+09	8.829E+08	2.097E+10	1.052E+09
29	32	2.030E+10	1.403E+09	8.829E+08	2.097E+10	1.052E+09
30	30	2.030E+10	1.403E+09	8.828E+08	2.097E+10	1.052E+09
31	28	2.030E+10	1.403E+09	8.828E+08	2.097E+10	1.052E+09
32	27	2.030E+10	1.403E+09	8.830E+08	2.097E+10	1.052E+09
33	26	2.030E+10	1.403E+09	8.827E+08	2.097E+10	1.051E+09
34	25	2.030E+10	1.403E+09	8.827E+08	2.097E+10	1.051E+09
35	23	2.030E+10	1.403E+09	8.826E+08	2.097E+10	1.051E+09
36	22	2.030E+10	1.403E+09	8.827E+08	2.097E+10	1.051E+09
37	20	2.030E+10	1.403E+09	8.826E+08	2.098E+10	1.051E+09
38	19	2.030E+10	1.403E+09	8.829E+08	2.098E+10	1.052E+09
39	18	2.030E+10	1.403E+09	8.820E+08	2.098E+10	1.052E+09
40	16	2.031E+10	1.403E+09	8.819E+08	2.098E+10	1.053E+09
41	13	2.031E+10	1.403E+09	8.843E+08	2.098E+10	1.054E+09
42	12	2.033E+10	1.404E+09	8.861E+08	2.099E+10	1.059E+09
43	8	2.036E+10	1.404E+09	8.828E+08	2.102E+10	1.065E+09
44	5	2.046E+10	1.404E+09	8.610E+08	2.104E+10	1.100E+09
45++	4	2.053E+10	1.405E+09	8.648E+08	2.125E+10	1.082E+09
46**	2	2.086E+10	1.407E+09	8.940E+08	2.172E+10	1.132E+09
47	1	2.708E+10	2.422E+09	1.300E+09	2.839E+10	1.542E+09

0-SE tree based on mean is marked with * and has 70 terminal nodes
0-SE tree based on median is marked with + and has 58 terminal nodes
Selected-SE tree based on mean using naive SE is marked with **
Selected-SE tree based on mean using bootstrap SE is marked with --
Selected-SE tree based on median and bootstrap SE is marked with ++
++ tree same as -- tree

Following tree is based on mean CV with naive SE estimate (**).

Structure of final tree. Each terminal node is marked with a T.

D-mean is weighted mean of INTP in the node
Cases fit give the number of cases used to fit node
MSE is residual sum of squares divided by number of cases in node

Node label	Total cases	Cases fit	Matrix rank	Node D-mean	Node MSE	Split variable	Interacting variable
1	39656	39656	1	2.008E+03	2.708E+10	PINCP	
2T	39063	39063	1	9.960E+02	3.026E+09	AGEP	
3T	593	593	1	6.491E+04	1.194E+12	-	

Number of terminal nodes of final tree: 2
Total number of nodes of final tree: 3

Regression tree:

Node 1: PINCP <= 253500.00
  Node 2: INTP-mean = 996.04400
Node 1: PINCP > 253500.00 or NA
  Node 3: INTP-mean = 64909.595

****

Predictor means below are weighted means of cases with no missing values.

WARNING: p-values below not adjusted for split search. For a bootstrap solution see:

1. Loh et al. (2016), "Identification of subgroups with differential treatment effects
for longitudinal and multiresponse variables", Statistics in Medicine, v.35, 4837-4855.

2. Loh et al. (2019), "Subgroups from regression trees with adjustment for prognostic
effects and post-selection inference", Statistics in Medicine, v.38, 545-557.

3. Loh and Zhou (2020), "The GUIDE approach to subgroup identification",
in "Design and Analysis of Subgroups with Biopharmaceutical Applications", Springer, pp.147-165.

Node 1: Intermediate node
A case goes into Node 2 if PINCP <= 253500.00
PINCP mean = 49327.164
Coefficients of least squares regression function:

Regressor	Coefficient	t-stat	p-value
Constant	2008.	15.87	0.000

INTP mean = 2007.62

Node 2: Terminal node
Coefficients of least squares regression functions:

Regressor	Coefficient	t-stat	p-value
Constant	996.0	23.35	0.2998E-14

INTP mean = 996.044

Node 3: Terminal node
Coefficients of least squares regression functions:

Regressor	Coefficient	t-stat	p-value
Constant	0.6491E+05	10.39	0.9992E-15

INTP mean = 64909.6

Proportion of variance (R-squared) explained by tree model: 0.2313

Observed and fitted values are stored in fit.txt
LaTeX code for tree is in code.tex
Elapsed time in seconds: 327.54

Output file of regression stepwise linear tree

Least squares regression tree
Predictions truncated at global min. and max. of D sample values
Pruning by cross-validation
Data description file: descreg.txt
Training sample file: data.txt
Missing value code: NA
Records in data file start on line 2
D variable is INTP
Piecewise forward and backward stepwise regression
F-to-enter and F-to-delete: 4.000 3.990
Using as many variables as needed
Number of records in data file: 46693
Length of longest entry in data file: 13
Missing values found in D variable
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Smallest and largest positive weights are 1.0000E+00 and 2.4080E+03

Summary information for training sample of size 39656 (excluding observations with non-positive weight or missing values in d, e, t, r or z variables)
d=dependent, b=split and fit cat variable using indicator variables,
c=split-only categorical, i=fit-only categorical (via indicators),
s=split-only numerical, n=split and fit numerical, f=fit-only numerical,
m=missing-value flag variable, p=periodic variable, w=weight

Column	Name		Minimum	Maximum	#Codes/ Levels/ Periods	#Missing
9	PWGTP	w	1.000	2408.		
10	AGEP	n	15.00	95.00		
13	COW	c			9	8906
23	FER	c			2	30173
24	GCL	c			2	7529
27	HIMRKS	c			3	
28	HINS1	c			2	
29	HINS2	c			2	
30	HINS3	c			2	
35	INTP	d	-1300.	0.2540E+06		
36	JWMNP	n	1.0000E+00	163.0		17020
38	JWTRNS	c			12	15212
40	MAR	c			5	
41	MARHD	c			2	10258
42	MARHM	c			2	10258
43	MARHT	c			3	10258
44	MARHW	c			2	10258

```
 45  MARHYP     n   1.9400E+03    2019.              10258
 57  NWAB       c                    3       672
 59  NWLA       c                    3       672
 60  NWLK       c                    3       672
 64  RELSHIPP   c                   19
 65  RETP       n    0.000       0.1420E+06
 66  SCH        c                    3
 68  SCHL       c                   24
 70  SEX        c                    2
 72  SSP        n    0.000       0.3690E+05
 73  WAGP       n    0.000       0.4760E+06
 74  WKHP       n   1.0000E+00    99.00             12232
 75  WKL        c                    3       672
 76  WKWN       n   1.0000E+00    52.00             12232
 77  WRK        c                    2      1386
 86  ESR        c                    5       672
 92  JWAP       n   1.0000E+00    285.0             17020
 93  JWDP       n   1.0000E+00    150.0             17020
 97  MSP        c                    6
101  OC         c                    2      1665
104  PERNP      n  -8.0000E+03   0.8310E+06             672
105  PINCP      n  -9300.        0.8590E+06
107  POVPIP     n   0.0000E+00    501.0              1126
111  PUBCOV     c                    2
119  RACNH      c                    2
124  RC         c                    2      1665
167  FINTP      c                    3

 Total  #cases w/   #missing
#cases    miss. D  ord. vals   #X-var   #N-var   #F-var   #S-var
 46693       7037      30071      244       13        0        0
#P-var   #M-var   #B-var   #C-var   #I-var
    0        0        0       29        0
```

Weight variable PWGTP in column: 9
Number of cases used for training: 39656
Number of split variables: 42
Number of cases excluded due to 0 weight or missing D: 7037

Missing values imputed with node means for fitting regression models in nodes
Pruning by v-fold cross-validation, with v = 10
Selected tree is based on mean of CV estimates
Number of SE's for pruned tree: 0.5000

Weighted error estimates used for pruning
Nodewise interaction tests on all variables
Fraction of cases used for splitting each node: 1.0000

Maximum number of split levels: 30
Minimum node sample size: 396
Top-ranked variables and chi-squared values at root node

1	0.1229E+05	AGEP
2	0.8019E+04	GCL
3	0.7139E+04	MARHT
4	0.6648E+04	MARHM
5	0.6544E+04	MARHD
6	0.6536E+04	MARHW
7	0.6485E+04	MAR
8	0.6445E+04	MSP
9	0.4211E+04	RELSHIPP
10	0.2454E+04	SCH
11	0.2284E+04	FER
12	0.1667E+04	MARHYP
13	0.1496E+04	HINS3
14	0.1369E+04	SSP
15	0.1332E+04	COW
16	0.1176E+04	NWLK
17	0.1124E+04	WKL
18	0.1085E+04	ESR
19	0.8609E+03	SCHL
20	0.8432E+03	NWAB
21	0.8410E+03	NWLA
22	0.7527E+03	RC
23	0.7269E+03	OC
24	0.7060E+03	WRK
25	0.6615E+03	WAGP
26	0.5368E+03	JWTRNS
27	0.5232E+03	RETP
28	0.4791E+03	PUBCOV
29	0.3834E+03	SEX
30	0.3782E+03	HINS2
31	0.3736E+03	HIMRKS
32	0.2847E+03	PERNP
33	0.2103E+03	WKHP
34	0.2035E+03	JWMNP
35	0.1701E+03	PINCP
36	0.1298E+03	JWAP
37	0.1219E+03	JWDP
38	0.1017E+03	HINS1
39	0.9338E+01	POVPIP

Size and CV MSE and SE of subtrees:

Tree	#Tnodes	Mean MSE	SE(Mean)	BSE(Mean)	Median MSE	BSE(Median)
1	64	1.761E+09	3.239E+08	2.596E+08	1.532E+09	2.617E+08

2	63	1.761E+09	3.239E+08	2.596E+08	1.532E+09	2.617E+08
3	62	1.761E+09	3.239E+08	2.596E+08	1.532E+09	2.617E+08
4	60	1.760E+09	3.239E+08	2.596E+08	1.530E+09	2.622E+08
5	59	1.760E+09	3.239E+08	2.596E+08	1.530E+09	2.622E+08
6	58	1.760E+09	3.239E+08	2.596E+08	1.530E+09	2.622E+08
7	57	1.760E+09	3.239E+08	2.596E+08	1.530E+09	2.622E+08
8	56	1.760E+09	3.239E+08	2.596E+08	1.530E+09	2.621E+08
9	55	1.761E+09	3.239E+08	2.595E+08	1.533E+09	2.615E+08
10	54	1.761E+09	3.239E+08	2.595E+08	1.533E+09	2.615E+08
11	53	1.761E+09	3.239E+08	2.595E+08	1.533E+09	2.615E+08
12	52	1.761E+09	3.239E+08	2.596E+08	1.533E+09	2.617E+08
13	51	1.760E+09	3.239E+08	2.596E+08	1.533E+09	2.612E+08
14	50	1.761E+09	3.239E+08	2.596E+08	1.533E+09	2.613E+08
15	49	1.761E+09	3.239E+08	2.596E+08	1.533E+09	2.615E+08
16	47	1.761E+09	3.239E+08	2.596E+08	1.533E+09	2.615E+08
17	46	1.760E+09	3.239E+08	2.596E+08	1.533E+09	2.619E+08
18	45	1.760E+09	3.239E+08	2.596E+08	1.533E+09	2.618E+08
19	44	1.760E+09	3.239E+08	2.596E+08	1.533E+09	2.618E+08
20	43	1.760E+09	3.239E+08	2.596E+08	1.533E+09	2.617E+08
21	41	1.760E+09	3.239E+08	2.597E+08	1.534E+09	2.618E+08
22	39	1.760E+09	3.239E+08	2.597E+08	1.534E+09	2.618E+08
23	38	1.760E+09	3.239E+08	2.597E+08	1.534E+09	2.622E+08
24	37	1.758E+09	3.239E+08	2.589E+08	1.532E+09	2.618E+08
25	36	1.758E+09	3.239E+08	2.589E+08	1.532E+09	2.617E+08
26	35	1.758E+09	3.239E+08	2.589E+08	1.532E+09	2.617E+08
27	33	1.758E+09	3.239E+08	2.589E+08	1.532E+09	2.617E+08
28	31	1.758E+09	3.239E+08	2.589E+08	1.532E+09	2.600E+08
29	30	1.757E+09	3.239E+08	2.589E+08	1.532E+09	2.581E+08
30	29	1.756E+09	3.239E+08	2.590E+08	1.531E+09	2.594E+08
31	28	1.757E+09	3.239E+08	2.591E+08	1.531E+09	2.602E+08
32	27	1.756E+09	3.239E+08	2.591E+08	1.531E+09	2.592E+08
33	26	1.756E+09	3.239E+08	2.591E+08	1.531E+09	2.592E+08
34	25	1.753E+09	3.239E+08	2.590E+08	1.531E+09	2.587E+08
35	24	1.752E+09	3.239E+08	2.592E+08	1.521E+09	2.608E+08
36	23	1.754E+09	3.242E+08	2.593E+08	1.522E+09	2.613E+08
37	21	1.753E+09	3.242E+08	2.593E+08	1.522E+09	2.619E+08
38	20	1.762E+09	3.244E+08	2.585E+08	1.528E+09	2.521E+08
39	19	1.768E+09	3.244E+08	2.579E+08	1.528E+09	2.546E+08
40	18	1.672E+09	2.302E+08	1.681E+08	1.528E+09	2.736E+08
41	17	1.680E+09	2.303E+08	1.657E+08	1.537E+09	2.747E+08
42	16	1.684E+09	2.304E+08	1.676E+08	1.537E+09	2.736E+08
43	15	1.705E+09	2.317E+08	1.752E+08	1.530E+09	3.126E+08
44	14	1.699E+09	2.315E+08	1.770E+08	1.535E+09	3.080E+08
45	12	1.690E+09	2.313E+08	1.792E+08	1.526E+09	3.134E+08
46	11	1.691E+09	2.313E+08	1.798E+08	1.526E+09	3.135E+08
47	9	1.700E+09	2.311E+08	1.793E+08	1.539E+09	3.120E+08

48*	8	1.549E+09	1.701E+08	1.390E+08	1.359E+09	2.338E+08
49	7	1.555E+09	1.620E+08	1.238E+08	1.444E+09	1.520E+08
50**	6	1.569E+09	1.628E+08	1.261E+08	1.444E+09	1.699E+08
51	4	1.802E+09	1.987E+08	2.300E+08	1.500E+09	2.394E+08
52	2	2.271E+09	2.113E+08	2.049E+08	2.171E+09	2.336E+08
53	1	2.682E+09	1.711E+08	1.806E+08	2.564E+09	1.432E+08

0-SE tree based on mean is marked with * and has 8 terminal nodes
0-SE tree based on median is marked with + and has 8 terminal nodes
Selected-SE tree based on mean using naive SE is marked with **
Selected-SE tree based on mean using bootstrap SE is marked with --
Selected-SE tree based on median and bootstrap SE is marked with ++
* tree same as + tree
** tree same as ++ tree
** tree same as -- tree
++ tree same as -- tree

Following tree is based on mean CV with naive SE estimate (**).

Structure of final tree. Each terminal node is marked with a T.

D-mean is weighted mean of INTP in the node
Cases fit give the number of cases used to fit node
MSE and R^2 are based on all cases in node

Node label	Total cases	Cases fit	Matrix rank	Node D-mean	Node MSE	Node R^2	Split variable	Other variables
1	39656	39656	12	2.008E+03	4.191E+09	0.8453	AGEP	
2T	2532	2532	3	4.538E+00	5.867E+06	0.0048	PINCP	
3	37124	37124	7	2.152E+03	2.318E+09	0.9198	PERNP	
6	13081	13081	8	3.800E+03	4.628E+09	0.8955	PINCP	
12T	12677	12677	9	1.601E+03	2.070E+09	0.4585	PINCP	
13T	404	404	4	7.223E+04	2.286E+10	0.9748	-	
7	24043	24043	8	1.425E+03	9.688E+08	0.9524	PUBCOV	
14T	3986	3986	8	2.414E+03	1.287E+09	0.9644	PINCP	
15	20057	20057	8	1.254E+03	8.934E+08	0.9480	PINCP	
30T	19570	19570	10	5.966E+02	5.364E+08	0.6816	PINCP	
31T	487	487	5	2.548E+04	6.682E+09	0.9884	-	

Number of terminal nodes of final tree: 6
Total number of nodes of final tree: 11
Second best split variable (based on curvature test) at root node is GCL

Regression tree:
For categorical variable splits, values not in training data go to the right

Node 1: AGEP <= 18.500000

```
  Node 2: INTP-mean = 4.5376387
Node 1: AGEP > 18.500000 or NA
  Node 3: PERNP <= 5250.0000
    Node 6: PINCP <= 93680.000
      Node 12: INTP-mean = 1601.4413
    Node 6: PINCP > 93680.000 or NA
      Node 13: INTP-mean = 72231.383
  Node 3: PERNP > 5250.0000 or NA
    Node 7: PUBCOV = "1"
      Node 14: INTP-mean = 2413.7009
    Node 7: PUBCOV /= "1"
      Node 15: PINCP <= 235375.00
        Node 30: INTP-mean = 596.62720
      Node 15: PINCP > 235375.00 or NA
        Node 31: INTP-mean = 25480.397

**************************************************************
```

Predictor means below are weighted means of cases with no missing values.
Regression coefficients are computed from the complete cases.

WARNING: p-values below not adjusted for split search. For a bootstrap solution see:

1. Loh et al. (2016), "Identification of subgroups with differential treatment effects
for longitudinal and multiresponse variables", Statistics in Medicine, v.35, 4837-4855.

2. Loh et al. (2019), "Subgroups from regression trees with adjustment for prognostic
effects and post-selection inference", Statistics in Medicine, v.38, 545-557.

3. Loh and Zhou (2020), "The GUIDE approach to subgroup identification",
in "Design and Analysis of Subgroups with Biopharmaceutical Applications", Springer, pp.147-165.

Node 1: Intermediate node
A case goes into Node 2 if AGEP <= 18.500000
AGEP mean = 46.273292
Coefficients of least squares regression function:

Regressor	Coefficient	t-stat	p-value	Minimum	Mean	Maximum
Constant	0.1308E+06	21.64	0.1033E-13			
AGEP	-104.6	-41.97	0.000	15.00	46.27	95.00
MARHYP	-64.09	-21.28	0.2109E-14	1940.	1994.	2019.
RETP	-0.8266	-243.1	0.6661E-15	0.000	3118.	0.1420E+06
SSP	-0.7754	-117.4	0.2220E-15	0.000	2985.	0.3690E+05
WAGP	-0.6854E-01	-38.89	0.1443E-14	0.000	0.3805E+05	0.4760E+06
WKHP	10.75	3.134	0.1727E-02	1.000	37.72	99.00
WKWN	22.05	6.693	0.2206E-10	1.000	46.29	52.00

JWAP -3.239 -2.957 0.3108E-02 1.000 101.5 285.0
PERNP -0.7704 -315.8 0.2220E-15 -8000. 0.4107E+05 0.8310E+06
PINCP 0.8408 458.1 0.4441E-15 -9300. 0.4933E+05 0.8590E+06
POVPIP 0.9711 4.088 0.4365E-04 0.000 361.3 501.0
INTP mean = 2007.62
Predicted values truncated at -1300.00 & 254000.

Node 2: Terminal node
Coefficients of least squares regression functions:
Regressor Coefficient t-stat p-value Minimum Mean Maximum
Constant 7.034 0.5615 0.5745
WKHP -0.5962 -0.8650 0.3871 1.000 17.57 99.00
PINCP 0.3065E-02 3.496 0.4809E-03 -4500. 2603. 0.7800E+05
INTP mean = 4.53764
Predicted values truncated at -1300.00 & 254000.

Node 3: Intermediate node
A case goes into Node 6 if PERNP <= 5250.0000
PERNP mean = 43052.800

Node 6: Intermediate node
A case goes into Node 12 if PINCP <= 93680.000
PINCP mean = 23531.801

Node 12: Terminal node
Coefficients of least squares regression functions:
Regressor Coefficient t-stat p-value Minimum Mean Maximum
Constant 0.2701E+05 4.164 0.3146E-04
AGEP 12.97 4.128 0.3683E-04 19.00 59.70 95.00
MARHYP -14.66 -4.537 0.5752E-05 1940. 1981. 2019.
RETP -0.4469 -78.42 0.1110E-15 0.000 6398. 0.8000E+05
SSP -0.4174 -55.90 0.3331E-15 0.000 8225. 0.3690E+05
WKHP 10.60 1.432 0.1520 1.000 20.25 99.00
PERNP -0.3671 -9.613 0.9326E-14 -8000. 450.3 5200.
PINCP 0.4469 96.37 0.000 -9300. 0.1875E+05 0.9366E+05
POVPIP 2.555 8.671 0.3331E-14 0.000 286.2 501.0
INTP mean = 1601.44
Predicted values truncated at -1300.00 & 254000.

Node 13: Terminal node
Coefficients of least squares regression functions:
Regressor Coefficient t-stat p-value Minimum Mean Maximum
Constant -0.2318E+05 -7.254 0.2119E-11
RETP -0.9192 -62.68 0.4441E-15 0.000 0.7601E+05 0.1420E+06
SSP -0.9987 -15.49 0.000 0.000 0.1786E+05 0.3690E+05
PINCP 1.062 84.95 0.4441E-15 0.9370E+05 0.1725E+06 0.4329E+06

INTP mean = 72231.4
Predicted values truncated at -1300.00 & 254000.

Node 7: Intermediate node
A case goes into Node 14 if PUBCOV = "1"
PUBCOV mode = "2"

Node 14: Terminal node
Coefficients of least squares regression functions:

Regressor	Coefficient	t-stat	p-value	Minimum	Mean	Maximum
Constant	0.1519E+05	1.607	0.1081			
MARHYP	-7.894	-1.670	0.9506E-01	1944.	1992.	2019.
RETP	-0.9561	-161.8	0.000	0.000	3579.	0.1420E+06
SSP	-0.9280	-108.1	0.000	0.000	5255.	0.3690E+05
WAGP	-0.6462E-02	-2.829	0.4692E-02	0.000	0.2985E+05	0.4760E+06
PERNP	-0.9580	-258.4	0.000	5300.	0.3550E+05	0.6200E+06
PINCP	0.9665	321.5	0.1110E-15	4700.	0.4785E+05	0.7360E+06
POVPIP	-2.646	-5.670	0.1529E-07	21.00	301.6	501.0

INTP mean = 2413.70
Predicted values truncated at -1300.00 & 254000.

Node 15: Intermediate node
A case goes into Node 30 if PINCP <= 235375.00
PINCP mean = 68607.586

Node 30: Terminal node
Coefficients of least squares regression functions:

Regressor	Coefficient	t-stat	p-value	Minimum	Mean	Maximum
Constant	8758.	2.721	0.6518E-02			
JWMNP	-2.888	-3.724	0.1964E-03	1.000	25.36	163.0
MARHYP	-4.668	-2.902	0.3713E-02	1963.	2001.	2019.
RETP	-0.6650	-140.4	0.7772E-15	0.000	529.6	0.1420E+06
SSP	-0.6676	-40.84	0.000	0.000	66.23	0.3000E+05
WAGP	-0.6889E-02	-4.550	0.5391E-05	0.000	0.5628E+05	0.2350E+06
WKWN	9.436	4.596	0.4336E-05	1.000	49.45	52.00
JWDP	0.9535	1.412	0.1580	1.000	50.78	150.0
PERNP	-0.6667	-176.7	0.000	5300.	0.5826E+05	0.2350E+06
PINCP	0.6758	202.2	0.000	4300.	0.5973E+05	0.2352E+06

INTP mean = 596.627
Predicted values truncated at -1300.00 & 254000.

Node 31: Terminal node
Coefficients of least squares regression functions:

Regressor	Coefficient	t-stat	p-value	Minimum	Mean	Maximum
Constant	-3630.	-3.144	0.1772E-02			
RETP	-0.9802	-53.39	0.4441E-15	0.000	2981.	0.1420E+06

SSP -0.9944 -0.7537 0.4514 0.000 4.218 0.1600E+05
PERNP -0.9906 -199.3 0.000 6500. 0.3665E+06 0.8310E+06
PINCP 0.9981 188.5 0.000 0.2355E+06 0.3959E+06 0.8590E+06
INTP mean = 25480.4
Predicted values truncated at -1300.00 & 254000.

Proportion of variance (R-squared) explained by tree model: 0.9504

Observed and fitted values are stored in fit.txt
LaTeX code for tree is in code.tex
Elapsed time in seconds: 1177.2

Output file of classification importance scoring

Classification tree
No pruning
Data description file: desc.txt
Training sample file: sample.txt
Missing value code: NA
Records in data file start on line 2
23 N variables changed to S
D variable is FINTP
Number of records in data file: 6000
Length of longest entry in data file: 13
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Number of classes: 2
Training sample class proportions of D variable FINTP:

Class	#Cases	Proportion
C	913	0.15216667
D	5087	0.84783333

Total #cases	#cases w/ miss. D	#missing ord. vals	#X-var	#N-var	#F-var	#S-var
6000	0	5938	182	0	0	23

#P-var	#M-var	#B-var	#C-var	#I-var
0	0	0	82	0

Number of cases used for training: 6000
Number of split variables: 105
Number of cases excluded due to 0 weight or missing D: 0

Importance scoring of variables
Simple node models
Estimated priors
Unit misclassification costs
Univariate split highest priority
Interaction splits 2nd priority; no linear splits
Split values for N and S variables based on exhaustive search
Maximum number of split levels: 4
Minimum node sample size: 60
New description file includes only selected variables
New description file name: descclass.txt
Starting 300 permutations to standardize means of importance scores
Finished permutations to standardize means of importance scores
95 and 99% thresholds for unadjusted importance scores = 35.763 46.094

Structure of final tree. Each terminal node is marked with a T.

Node cost is node misclassification cost divided by number of training cases

Node label	Total cases	Train cases	Predicted class	Node cost	Split variables	Interacting variable
1	6000	6000	D	1.522E-01	INTP	
2T	913	913	C	1.413E-04	-	
3T	5087	5087	D	2.536E-05	-	

Number of terminal nodes of final tree: 2
Total number of nodes of final tree: 3
Second best split variable (based on curvature test) at root node is WRK

Classification tree:

Node 1: INTP = NA
  Node 2: C
Node 1: INTP /= NA
  Node 3: D

****

Predictor means below are means of cases with no missing values.

Node 1: Intermediate node
 A case goes into Node 2 if INTP = NA
 INTP mean = 2456.1293

Class	Number	Posterior
C	913	0.1522E+00
D	5087	0.8478E+00

 Number of training cases misclassified = 913
 Predicted class is D

Node 2: Terminal node

Class	Number	Posterior
C	913	0.9999E+00
D	0	0.1413E-03

 Number of training cases misclassified = 0
 Predicted class is C

Node 3: Terminal node

Class	Number	Posterior
C	0	0.2536E-04
D	5087	0.1000E+01

 Number of training cases misclassified = 0
 Predicted class is D

****

Variables used for splitting:
INTP

Number of terminal nodes: 2

Scaled importance scores of predictor variables

Score	Rank	Variable
1.700E+02	1.00	INTP
4.585E+01	2.00	WRK
3.115E+01	3.00	ANC
2.938E+00	4.00	RELSHIPP
2.621E+00	5.00	POVPIP
1.781E+00	6.00	DDRS
1.769E+00	7.00	NWAB
1.746E+00	8.00	NWLA
1.739E+00	9.00	NWLK
1.733E+00	10.00	RC
1.718E+00	11.00	OC
1.687E+00	12.00	MSP
1.594E+00	13.00	AGEP
1.414E+00	14.00	PERNP
1.356E+00	15.00	WKL
1.281E+00	16.00	MAR
--------- variables above this line are highly important -----		
1.170E+00	17.00	SCHL
1.116E+00	18.00	WKHP
1.114E+00	19.00	DPHY
1.086E+00	20.00	HINS3
1.039E+00	21.00	WKWN
1.038E+00	22.00	COW
--------- variables below this line are unimportant ----------		
9.619E-01	23.00	SSP
9.043E-01	24.00	WAGP
8.965E-01	25.00	DIS
8.923E-01	26.00	DOUT
8.586E-01	27.00	HINS1
7.918E-01	28.00	PUBCOV
7.014E-01	29.00	ESR
6.781E-01	30.00	JWMNP
6.674E-01	31.00	JWRIP
6.331E-01	32.00	JWDP
6.233E-01	33.00	RAC1P
5.746E-01	34.00	JWAP
5.219E-01	35.00	PRIVCOV
5.158E-01	36.00	JWTRNS
5.124E-01	37.00	PINCP

5.071E-01	38.00	MARHYP
4.742E-01	39.00	DREM
4.646E-01	40.00	RACBLK
4.088E-01	41.00	MLPJ
3.799E-01	42.00	POWPUMA
3.787E-01	43.00	PAOC
3.742E-01	44.00	RACWHT
3.567E-01	45.00	MLPCD
3.555E-01	46.00	MLPB
3.372E-01	47.00	VPS
3.181E-01	48.00	MLPI
2.850E-01	49.00	FER
2.798E-01	50.00	MIG
2.754E-01	51.00	MLPFG
2.749E-01	52.00	DRATX
2.660E-01	53.00	RACNUM
2.592E-01	54.00	MLPA
2.587E-01	55.00	MLPK
2.487E-01	56.00	HICOV
2.376E-01	57.00	MLPH
2.303E-01	58.00	MIL
2.235E-01	59.00	MARHW
2.153E-01	60.00	MLPE
1.781E-01	61.00	RACNH
1.588E-01	62.00	HINS4
1.555E-01	63.00	PWGTP
1.554E-01	64.00	HINS2
1.515E-01	65.00	SEMP
1.366E-01	66.00	DEAR
1.325E-01	67.00	HINS6
1.215E-01	68.00	DEYE
1.167E-01	69.00	HIMRKS
1.114E-01	70.00	QTRBIR
1.024E-01	71.00	DRAT
9.515E-02	72.00	RETP
6.770E-02	73.00	RACPI
6.249E-02	74.00	HINS7
6.188E-02	75.00	SSIP
6.024E-02	76.00	MARHD
5.894E-02	77.00	SCHG
5.880E-02	78.00	OIP
5.802E-02	79.00	NATIVITY
5.351E-02	80.00	RACSOR
5.320E-02	81.00	MIGPUMA
5.127E-02	82.00	LANX
4.961E-02	83.00	MARHM

```
4.106E-02   84.00  SEX
3.857E-02   85.00  YOEP
3.561E-02   86.00  NWAV
3.506E-02   87.00  MARHT
2.978E-02   88.00  GCR
2.585E-02   89.00  ESP
2.168E-02   90.00  CIT
1.851E-02   91.00  HISP
1.215E-02   92.00  HINS5
1.178E-02   93.00  NWRE
1.166E-02   94.00  WAOB
8.668E-03   95.00  ENG
7.102E-03   96.00  GCM
7.024E-03   97.00  GCL
6.949E-03   98.00  SFR
6.754E-03   99.00  SFN
6.373E-03  100.00  NOP
5.373E-03  101.00  CITWP
4.266E-03  102.00  RACAIAN
4.033E-03  103.00  SCH
3.827E-03  104.00  RACASN
6.763E-04  105.00  PAP

Variables with scores above 1.23 are highly important
Variables with scores between 1.0 and 1.23 are likely important
Variables with scores below 1.0 are unimportant
No. highly important, likely important, and unimportant split variables: 16, 6, 83
LaTeX code for tree is in descclass.txt
Importance scores are stored in imp.txt
Description file with selected variables in descclass.txt
Elapsed time in seconds: 828.40
```

Output file of regression importance scoring

```
Least squares regression tree
No pruning
Data description file: desc.txt
Training sample file: sample.txt
Missing value code: NA
Records in data file start on line 2
21 N variables changed to S
D variable is INTP
Piecewise constant model
Number of records in data file: 6000
Length of longest entry in data file: 13
Missing values found in D variable
Missing values found among categorical variables
Separate categories will be created for missing categorical variables
Missing values found among non-categorical variables
Smallest and largest positive weights are  2.0000E+00 and 2.4080E+03

    Total  #cases w/   #missing
   #cases    miss. D  ord. vals   #X-var   #N-var   #F-var   #S-var
     6000        913       5938      182        0        0       21
   #P-var   #M-var   #B-var   #C-var   #I-var
        0        0        0       83        0
Weight variable PWGTP in column: 9
Number of cases used for training: 5087
Number of split variables: 104
Number of cases excluded due to 0 weight or missing D: 913

Importance scoring of variables
Nodewise interaction tests on all variables
Split values for N and S variables based on exhaustive search
Maximum number of split levels: 4
Minimum node sample size: 50
New description file includes only selected variables
New description file name: descreg.txt
Starting 300 permutations to standardize means of importance scores
Finished permutations to standardize means of importance scores
95 and 99% thresholds for unadjusted importance scores =   54.867  68.647

Structure of final tree. Each terminal node is marked with a T.

D-mean is weighted mean of INTP in the node
Cases fit give the number of cases used to fit node
MSE is residual sum of squares divided by number of cases in node
      Node    Total    Cases Matrix   Node      Node      Split          Interacting
      label   cases      fit  rank    D-mean    MSE       variable       variable
```

1	5087	5087	1	1.812E+03	1.995E+10	AGEP
2	3835	3835	1	9.855E+02	8.402E+09	AGEP
4	2375	2375	1	3.599E+02	5.500E+09	PINCP
8	2325	2325	1	1.506E+02	2.298E+08	PINCP
16T	2202	2202	1	9.421E+01	1.556E+08	SCHL
17T	123	123	1	1.097E+03	1.452E+09	MARHYP
9T	50	50	1	1.006E+04	2.441E+11	-
5	1460	1460	1	2.322E+03	1.290E+10	PINCP
10	1409	1409	1	1.574E+03	5.369E+09	SEX
20T	681	681	1	2.255E+03	8.123E+09	HIMRKS
21T	728	728	1	9.186E+02	2.726E+09	ESR
11T	51	51	1	1.703E+04	1.965E+11	-
3	1252	1252	1	5.366E+03	5.417E+10	PINCP
6	1202	1202	1	2.491E+03	5.667E+09	PINCP
12	1105	1105	1	1.495E+03	2.103E+09	PINCP
24T	815	815	1	6.258E+02	3.922E+08	SCHL
25T	290	290	1	3.476E+03	6.445E+09	RETP
13T	97	97	1	1.131E+04	3.845E+10	-
7T	50	50	1	6.421E+04	9.168E+11	-

Number of terminal nodes of final tree: 10
Total number of nodes of final tree: 19
Second best split variable (based on curvature test) at root node is HINS3

Regression tree:
For categorical variable splits, values not in training data go to the right

```
Node 1: AGEP <= 64.500000
  Node 2: AGEP <= 49.500000
    Node 4: PINCP <= 172500.00
      Node 8: PINCP <= 99650.000
        Node 16: INTP-mean = 94.209927
      Node 8: PINCP > 99650.000 or NA
        Node 17: INTP-mean = 1097.4600
    Node 4: PINCP > 172500.00 or NA
      Node 9: INTP-mean = 10062.402
  Node 2: AGEP > 49.500000 or NA
    Node 5: PINCP <= 173000.00
      Node 10: SEX = "1"
        Node 20: INTP-mean = 2255.1356
      Node 10: SEX /= "1"
        Node 21: INTP-mean = 918.61540
    Node 5: PINCP > 173000.00 or NA
      Node 11: INTP-mean = 17030.068
Node 1: AGEP > 64.500000 or NA
  Node 3: PINCP <= 131950.00
```

Node 6: PINCP <= 75800.000
Node 12: PINCP <= 39900.000
Node 24: INTP-mean = 625.75482
Node 12: PINCP > 39900.000 or NA
Node 25: INTP-mean = 3476.1407
Node 6: PINCP > 75800.000 or NA
Node 13: INTP-mean = 11312.176
Node 3: PINCP > 131950.00 or NA
Node 7: INTP-mean = 64212.913

****

Predictor means below are weighted means of cases with no missing values.

WARNING: p-values below not adjusted for split search. For a bootstrap solution see:

1. Loh et al. (2016), "Identification of subgroups with differential treatment effects for longitudinal and multiresponse variables", Statistics in Medicine, v.35, 4837-4855.

2. Loh et al. (2019), "Subgroups from regression trees with adjustment for prognostic effects and post-selection inference", Statistics in Medicine, v.38, 545-557.

3. Loh and Zhou (2020), "The GUIDE approach to subgroup identification", in "Design and Analysis of Subgroups with Biopharmaceutical Applications", Springer, pp.147-165.

Node 1: Intermediate node
A case goes into Node 2 if AGEP <= 64.500000
AGEP mean = 46.204408
Coefficients of least squares regression function:
Regressor Coefficient t-stat p-value
Constant 1812. 5.882 0.4316E-08
INTP mean = 1811.99

Node 2: Intermediate node
A case goes into Node 4 if AGEP <= 49.500000
AGEP mean = 39.760079

Node 4: Intermediate node
A case goes into Node 8 if PINCP <= 172500.00
PINCP mean = 42801.224

Node 8: Intermediate node
A case goes into Node 16 if PINCP <= 99650.000
PINCP mean = 36653.000

Node 16: Terminal node
Coefficients of least squares regression functions:
Regressor Coefficient t-stat p-value
Constant 94.21 2.447 0.1448E-01
INTP mean = 94.2099

Node 17: Terminal node
Coefficients of least squares regression functions:
Regressor Coefficient t-stat p-value
Constant 1097. 2.712 0.7662E-02
INTP mean = 1097.46

Node 9: Terminal node
Coefficients of least squares regression functions:
Regressor Coefficient t-stat p-value
Constant 0.1006E+05 1.241 0.2205
INTP mean = 10062.4

Node 5: Intermediate node
A case goes into Node 10 if PINCP <= 173000.00
PINCP mean = 62129.871

Node 10: Intermediate node
A case goes into Node 20 if SEX = "1"
SEX mode = "2"

Node 20: Terminal node
Coefficients of least squares regression functions:
Regressor Coefficient t-stat p-value
Constant 2255. 4.015 0.6598E-04
INTP mean = 2255.14

Node 21: Terminal node
Coefficients of least squares regression functions:
Regressor Coefficient t-stat p-value
Constant 918.6 3.009 0.2709E-02
INTP mean = 918.615

Node 11: Terminal node
Coefficients of least squares regression functions:
Regressor Coefficient t-stat p-value
Constant 0.1703E+05 2.176 0.3429E-01
INTP mean = 17030.1

Node 3: Intermediate node
A case goes into Node 6 if PINCP <= 131950.00

PINCP mean = 46067.187

Node 6: Intermediate node
A case goes into Node 12 if PINCP <= 75800.000
PINCP mean = 36158.170

Node 12: Intermediate node
A case goes into Node 24 if PINCP <= 39900.000
PINCP mean = 29093.810

Node 24: Terminal node
Coefficients of least squares regression functions:

Regressor	Coefficient	t-stat	p-value
Constant	625.8	5.179	0.2811E-06

INTP mean = 625.755

Node 25: Terminal node
Coefficients of least squares regression functions:

Regressor	Coefficient	t-stat	p-value
Constant	3476.	4.651	0.5037E-05

INTP mean = 3476.14

Node 13: Terminal node
Coefficients of least squares regression functions:

Regressor	Coefficient	t-stat	p-value
Constant	0.1131E+05	3.992	0.1284E-03

INTP mean = 11312.2

Node 7: Terminal node
Coefficients of least squares regression functions:

Regressor	Coefficient	t-stat	p-value
Constant	0.6421E+05	3.445	0.1180E-02

INTP mean = 64212.9

****
Variables used for splitting or fitting:
AGEP
PINCP
SEX

Number of terminal nodes: 10

Scaled importance scores of predictor variables

Score	Rank	Variable
7.569E+00	1.00	PINCP
7.131E+00	2.00	AGEP

4.354E+00 3.00 HINS3
3.865E+00 4.00 RETP
3.833E+00 5.00 MARHYP
3.635E+00 6.00 SSP
3.205E+00 7.00 POVPIP
2.784E+00 8.00 HINS2
2.695E+00 9.00 COW
2.671E+00 10.00 MAR
2.572E+00 11.00 MSP
2.553E+00 12.00 FER
2.464E+00 13.00 WAGP
2.422E+00 14.00 HIMRKS
2.243E+00 15.00 SCHL
2.089E+00 16.00 RELSHIPP
2.078E+00 17.00 PUBCOV
2.029E+00 18.00 GCL
1.817E+00 19.00 PERNP
1.636E+00 20.00 MARHW
1.570E+00 21.00 WKHP
1.530E+00 22.00 WRK
1.459E+00 23.00 NWLK
1.459E+00 24.00 WKWN
1.450E+00 25.00 MARHT
1.443E+00 26.00 MARHM
1.421E+00 27.00 MARHD
1.412E+00 28.00 ESR
1.361E+00 29.00 NWAB
1.339E+00 30.00 NWLA
--------- variables above this line are highly important -----
1.329E+00 31.00 RC
1.288E+00 32.00 JWAP
1.267E+00 33.00 WKL
1.263E+00 34.50 RACNH
1.263E+00 34.50 FINTP
1.235E+00 36.00 JWMNP
1.218E+00 37.00 JWDP
1.184E+00 38.00 HINS1
1.141E+00 39.00 OC
1.058E+00 40.00 SEX
1.049E+00 41.00 SCH
1.001E+00 42.00 JWTRNS
--------- variables below this line are unimportant ----------
9.991E-01 43.00 JWRIP
9.453E-01 44.00 ESP
9.167E-01 45.00 POWPUMA
8.878E-01 46.00 SEMP

8.355E-01	47.00	NOP
8.294E-01	48.00	HINS4
7.549E-01	49.00	MIG
7.537E-01	50.00	SCHG
7.068E-01	51.00	RACPI
7.042E-01	52.00	PAOC
6.431E-01	53.00	VPS
6.019E-01	54.00	RACWHT
5.110E-01	55.00	ANC
4.707E-01	56.00	DOUT
3.806E-01	57.00	MIL
3.572E-01	58.00	MLPFG
3.552E-01	59.00	LANX
3.179E-01	60.00	RAC1P
3.016E-01	61.00	RACBLK
2.910E-01	62.00	DRATX
2.894E-01	63.00	MLPJ
2.831E-01	64.00	SFN
2.719E-01	65.00	DPHY
2.456E-01	66.00	DIS
2.440E-01	67.00	NWRE
2.397E-01	68.00	DREM
2.344E-01	69.00	SSIP
2.304E-01	70.00	MLPA
2.184E-01	71.00	RACAIAN
1.711E-01	72.00	NWAV
1.709E-01	73.00	MLPI
1.691E-01	74.00	WAOB
1.672E-01	75.00	PRIVCOV
1.639E-01	76.00	ENG
1.563E-01	77.00	DDRS
1.551E-01	78.00	MLPB
1.528E-01	79.00	PAP
1.528E-01	80.00	MLPK
1.496E-01	81.00	MLPCD
1.491E-01	82.00	HINS7
1.412E-01	83.00	MLPE
1.393E-01	84.00	NATIVITY
1.341E-01	85.00	HICOV
1.338E-01	86.00	MLPH
1.230E-01	87.00	YOEP
1.170E-01	88.00	RACASN
1.142E-01	89.00	CIT
1.056E-01	90.00	RACSOR
8.888E-02	91.00	CITWP
8.364E-02	92.00	DEAR

7.947E-02	93.00	OIP
7.389E-02	94.00	HINS6
6.745E-02	95.00	HINS5
6.408E-02	96.00	GCR
5.199E-02	97.00	RACNUM
4.801E-02	98.00	SFR
4.096E-02	99.00	DEYE
3.887E-02	100.00	MIGPUMA
3.243E-02	101.00	QTRBIR
2.525E-02	102.00	GCM
2.368E-02	103.00	HISP
8.124E-03	104.00	DRAT

Variables with scores above 1.33 are highly important
Variables with scores between 1.0 and 1.33 are likely important
Variables with scores below 1.0 are unimportant
No. highly important, likely important, and unimportant split variables: 30, 12, 62
LaTeX code for tree is in descreg.txt
Importance scores are stored in imp.txt
Description file with selected variables in descreg.txt
Elapsed time in seconds: 871.01

Description file for classification importance scoring

sample.txt
NA
2
1 RT x
2 SERIALNO x
3 DIVISION x
4 SPORDER x
5 PUMA x
6 REGION x
7 ST x
8 ADJINC x
9 PWGTP w
10 AGEP n
11 CIT c
12 CITWP n
13 COW c
14 DDRS c
15 DEAR c
16 DEYE c
17 DOUT c
18 DPHY c
19 DRAT c
20 DRATX c
21 DREM c
22 ENG c
23 FER c
24 GCL c
25 GCM c
26 GCR c
27 HIMRKS c
28 HINS1 c
29 HINS2 c
30 HINS3 c
31 HINS4 c
32 HINS5 c
33 HINS6 c
34 HINS7 c
35 INTP n
36 JWMNP n
37 JWRIP n
38 JWTRNS c
39 LANX c
40 MAR c
41 MARHD c
42 MARHM c

43 MARHT c
44 MARHW c
45 MARHYP n
46 MIG c
47 MIL c
48 MLPA c
49 MLPB c
50 MLPCD c
51 MLPE c
52 MLPFG c
53 MLPH c
54 MLPI c
55 MLPJ c
56 MLPK c
57 NWAB c
58 NWAV c
59 NWLA c
60 NWLK c
61 NWRE c
62 OIP n
63 PAP n
64 RELSHIPP c
65 RETP n
66 SCH c
67 SCHG c
68 SCHL c
69 SEMP n
70 SEX c
71 SSIP n
72 SSP n
73 WAGP n
74 WKHP n
75 WKL c
76 WKWN n
77 WRK c
78 YOEP n
79 ANC c
80 ANC1P x
81 ANC2P x
82 DECADE x
83 DIS c
84 DRIVESP x
85 ESP c
86 ESR c
87 FOD1P x
88 FOD2P x

89 HICOV c
90 HISP c
91 INDP x
92 JWAP n
93 JWDP n
94 LANP x
95 MIGPUMA c
96 MIGSP x
97 MSP c
98 NAICSP x
99 NATIVITY c
100 NOP c
101 OC c
102 OCCP x
103 PAOC c
104 PERNP n
105 PINCP n
106 POBP x
107 POVPIP n
108 POWPUMA c
109 POWSP x
110 PRIVCOV c
111 PUBCOV c
112 QTRBIR c
113 RAC1P c
114 RAC2P x
115 RAC3P x
116 RACAIAN c
117 RACASN c
118 RACBLK c
119 RACNH c
120 RACNUM n
121 RACPI c
122 RACSOR c
123 RACWHT c
124 RC c
125 SCIENGP x
126 SCIENGRLP x
127 SFN c
128 SFR c
129 SOCP x
130 VPS c
131 WAOB c
132 FAGEP x
133 FANCP x
134 FCITP x

135 FCITWP x
136 FCOWP x
137 FDDRSP x
138 FDEARP x
139 FDEYEP x
140 FDISP x
141 FDOUTP x
142 FDPHYP x
143 FDRATP x
144 FDRATXP x
145 FDREMP x
146 FENGP x
147 FESRP x
148 FFERP x
149 FFODP x
150 FGCLP x
151 FGCMP x
152 FGCRP x
153 FHICOVP x
154 FHIMRKSP x
155 FHINS1P x
156 FHINS2P x
157 FHINS3C x
158 FHINS3P x
159 FHINS4C x
160 FHINS4P x
161 FHINS5C x
162 FHINS5P x
163 FHINS6P x
164 FHINS7P x
165 FHISP x
166 FINDP x
167 FINTP d
168 FJWDP x
169 FJWMNP x
170 FJWRIP x
171 FJWTRNSP x
172 FLANP x
173 FLANXP x
174 FMARP x
175 FMARHDP x
176 FMARHMP x
177 FMARHTP x
178 FMARHWP x
179 FMARHYP x
180 FMIGP x

181 FMIGSP x
182 FMILPP x
183 FMILSP x
184 FOCCP x
185 FOIP x
186 FPAP x
187 FPERNP x
188 FPINCP x
189 FPOBP x
190 FPOWSP x
191 FPRIVCOVP x
192 FPUBCOVP x
193 FRACP x
194 FRELSHIPP x
195 FRETP x
196 FSCHGP x
197 FSCHLP x
198 FSCHP x
199 FSEMP x
200 FSEXP x
201 FSSIP x
202 FSSP x
203 FWAGP x
204 FWKHP x
205 FWKLP x
206 FWKWNP x
207 FWRKP x
208 FYOEP x
209 PWGTP1 x
210 PWGTP2 x
211 PWGTP3 x
212 PWGTP4 x
213 PWGTP5 x
214 PWGTP6 x
215 PWGTP7 x
216 PWGTP8 x
217 PWGTP9 x
218 PWGTP10 x
219 PWGTP11 x
220 PWGTP12 x
221 PWGTP13 x
222 PWGTP14 x
223 PWGTP15 x
224 PWGTP16 x
225 PWGTP17 x
226 PWGTP18 x

227 PWGTP19 x
228 PWGTP20 x
229 PWGTP21 x
230 PWGTP22 x
231 PWGTP23 x
232 PWGTP24 x
233 PWGTP25 x
234 PWGTP26 x
235 PWGTP27 x
236 PWGTP28 x
237 PWGTP29 x
238 PWGTP30 x
239 PWGTP31 x
240 PWGTP32 x
241 PWGTP33 x
242 PWGTP34 x
243 PWGTP35 x
244 PWGTP36 x
245 PWGTP37 x
246 PWGTP38 x
247 PWGTP39 x
248 PWGTP40 x
249 PWGTP41 x
250 PWGTP42 x
251 PWGTP43 x
252 PWGTP44 x
253 PWGTP45 x
254 PWGTP46 x
255 PWGTP47 x
256 PWGTP48 x
257 PWGTP49 x
258 PWGTP50 x
259 PWGTP51 x
260 PWGTP52 x
261 PWGTP53 x
262 PWGTP54 x
263 PWGTP55 x
264 PWGTP56 x
265 PWGTP57 x
266 PWGTP58 x
267 PWGTP59 x
268 PWGTP60 x
269 PWGTP61 x
270 PWGTP62 x
271 PWGTP63 x
272 PWGTP64 x

273 PWGTP65 x
274 PWGTP66 x
275 PWGTP67 x
276 PWGTP68 x
277 PWGTP69 x
278 PWGTP70 x
279 PWGTP71 x
280 PWGTP72 x
281 PWGTP73 x
282 PWGTP74 x
283 PWGTP75 x
284 PWGTP76 x
285 PWGTP77 x
286 PWGTP78 x
287 PWGTP79 x
288 PWGTP80 x

Description file for regression importance scoring

sample.txt
NA
2
1 RT x
2 SERIALNO x
3 DIVISION x
4 SPORDER x
5 PUMA x
6 REGION x
7 ST x
8 ADJINC x
9 PWGTP w
10 AGEP n
11 CIT c
12 CITWP n
13 COW c
14 DDRS c
15 DEAR c
16 DEYE c
17 DOUT c
18 DPHY c
19 DRAT c
20 DRATX c
21 DREM c
22 ENG c
23 FER c
24 GCL c
25 GCM c
26 GCR c
27 HIMRKS c
28 HINS1 c
29 HINS2 c
30 HINS3 c
31 HINS4 c
32 HINS5 c
33 HINS6 c
34 HINS7 c
35 INTP d
36 JWMNP n
37 JWRIP n
38 JWTRNS c
39 LANX c
40 MAR c
41 MARHD c
42 MARHM c

43 MARHT c
44 MARHW c
45 MARHYP n
46 MIG c
47 MIL c
48 MLPA c
49 MLPB c
50 MLPCD c
51 MLPE c
52 MLPFG c
53 MLPH c
54 MLPI c
55 MLPJ c
56 MLPK c
57 NWAB c
58 NWAV c
59 NWLA c
60 NWLK c
61 NWRE c
62 OIP n
63 PAP n
64 RELSHIPP c
65 RETP n
66 SCH c
67 SCHG c
68 SCHL c
69 SEMP n
70 SEX c
71 SSIP n
72 SSP n
73 WAGP n
74 WKHP n
75 WKL c
76 WKWN n
77 WRK c
78 YOEP n
79 ANC c
80 ANC1P x
81 ANC2P x
82 DECADE x
83 DIS c
84 DRIVESP x
85 ESP c
86 ESR c
87 FOD1P x
88 FOD2P x

89 HICOV c
90 HISP c
91 INDP x
92 JWAP n
93 JWDP n
94 LANP x
95 MIGPUMA c
96 MIGSP x
97 MSP c
98 NAICSP x
99 NATIVITY c
100 NOP c
101 OC c
102 OCCP x
103 PAOC c
104 PERNP n
105 PINCP n
106 POBP x
107 POVPIP n
108 POWPUMA c
109 POWSP x
110 PRIVCOV c
111 PUBCOV c
112 QTRBIR c
113 RAC1P c
114 RAC2P x
115 RAC3P x
116 RACAIAN c
117 RACASN c
118 RACBLK c
119 RACNH c
120 RACNUM n
121 RACPI c
122 RACSOR c
123 RACWHT c
124 RC c
125 SCIENGP x
126 SCIENGRLP x
127 SFN c
128 SFR c
129 SOCP x
130 VPS c
131 WAOB c
132 FAGEP x
133 FANCP x
134 FCITP x

135 FCITWP x
136 FCOWP x
137 FDDRSP x
138 FDEARP x
139 FDEYEP x
140 FDISP x
141 FDOUTP x
142 FDPHYP x
143 FDRATP x
144 FDRATXP x
145 FDREMP x
146 FENGP x
147 FESRP x
148 FFERP x
149 FFODP x
150 FGCLP x
151 FGCMP x
152 FGCRP x
153 FHICOVP x
154 FHIMRKSP x
155 FHINS1P x
156 FHINS2P x
157 FHINS3C x
158 FHINS3P x
159 FHINS4C x
160 FHINS4P x
161 FHINS5C x
162 FHINS5P x
163 FHINS6P x
164 FHINS7P x
165 FHISP x
166 FINDP x
167 FINTP c
168 FJWDP x
169 FJWMNP x
170 FJWRIP x
171 FJWTRNSP x
172 FLANP x
173 FLANXP x
174 FMARP x
175 FMARHDP x
176 FMARHMP x
177 FMARHTP x
178 FMARHWP x
179 FMARHYP x
180 FMIGP x

181 FMIGSP x
182 FMILPP x
183 FMILSP x
184 FOCCP x
185 FOIP x
186 FPAP x
187 FPERNP x
188 FPINCP x
189 FPOBP x
190 FPOWSP x
191 FPRIVCOVP x
192 FPUBCOVP x
193 FRACP x
194 FRELSHIPP x
195 FRETP x
196 FSCHGP x
197 FSCHLP x
198 FSCHP x
199 FSEMP x
200 FSEXP x
201 FSSIP x
202 FSSP x
203 FWAGP x
204 FWKHP x
205 FWKLP x
206 FWKWNP x
207 FWRKP x
208 FYOEP x
209 PWGTP1 x
210 PWGTP2 x
211 PWGTP3 x
212 PWGTP4 x
213 PWGTP5 x
214 PWGTP6 x
215 PWGTP7 x
216 PWGTP8 x
217 PWGTP9 x
218 PWGTP10 x
219 PWGTP11 x
220 PWGTP12 x
221 PWGTP13 x
222 PWGTP14 x
223 PWGTP15 x
224 PWGTP16 x
225 PWGTP17 x
226 PWGTP18 x

227 PWGTP19 x
228 PWGTP20 x
229 PWGTP21 x
230 PWGTP22 x
231 PWGTP23 x
232 PWGTP24 x
233 PWGTP25 x
234 PWGTP26 x
235 PWGTP27 x
236 PWGTP28 x
237 PWGTP29 x
238 PWGTP30 x
239 PWGTP31 x
240 PWGTP32 x
241 PWGTP33 x
242 PWGTP34 x
243 PWGTP35 x
244 PWGTP36 x
245 PWGTP37 x
246 PWGTP38 x
247 PWGTP39 x
248 PWGTP40 x
249 PWGTP41 x
250 PWGTP42 x
251 PWGTP43 x
252 PWGTP44 x
253 PWGTP45 x
254 PWGTP46 x
255 PWGTP47 x
256 PWGTP48 x
257 PWGTP49 x
258 PWGTP50 x
259 PWGTP51 x
260 PWGTP52 x
261 PWGTP53 x
262 PWGTP54 x
263 PWGTP55 x
264 PWGTP56 x
265 PWGTP57 x
266 PWGTP58 x
267 PWGTP59 x
268 PWGTP60 x
269 PWGTP61 x
270 PWGTP62 x
271 PWGTP63 x
272 PWGTP64 x

273 PWGTP65 x
274 PWGTP66 x
275 PWGTP67 x
276 PWGTP68 x
277 PWGTP69 x
278 PWGTP70 x
279 PWGTP71 x
280 PWGTP72 x
281 PWGTP73 x
282 PWGTP74 x
283 PWGTP75 x
284 PWGTP76 x
285 PWGTP77 x
286 PWGTP78 x
287 PWGTP79 x
288 PWGTP80 x

Description file for classification tree and forest after excluding unimportant variables

```
"data.txt"
 "NA"
 2
 1 RT x
 2 SERIALNO x
 3 DIVISION x
 4 SPORDER x
 5 PUMA x
 6 REGION x
 7 ST x
 8 ADJINC x
 9 PWGTP x
 10 AGEP n
 11 CIT x
 12 CITWP x
 13 COW c
 14 DDRS c
 15 DEAR x
 16 DEYE x
 17 DOUT x
 18 DPHY c
 19 DRAT x
 20 DRATX x
 21 DREM x
 22 ENG x
 23 FER x
 24 GCL x
 25 GCM x
 26 GCR x
 27 HIMRKS x
 28 HINS1 x
 29 HINS2 x
 30 HINS3 c
 31 HINS4 x
 32 HINS5 x
 33 HINS6 x
 34 HINS7 x
 35 INTP x
 36 JWMNP x
 37 JWRIP x
 38 JWTRNS x
 39 LANX x
 40 MAR c
 41 MARHD x
 42 MARHM x
```

43 MARHT x
44 MARHW x
45 MARHYP x
46 MIG x
47 MIL x
48 MLPA x
49 MLPB x
50 MLPCD x
51 MLPE x
52 MLPFG x
53 MLPH x
54 MLPI x
55 MLPJ x
56 MLPK x
57 NWAB c
58 NWAV x
59 NWLA c
60 NWLK c
61 NWRE x
62 OIP x
63 PAP x
64 RELSHIPP c
65 RETP x
66 SCH x
67 SCHG x
68 SCHL c
69 SEMP x
70 SEX x
71 SSIP x
72 SSP x
73 WAGP x
74 WKHP n
75 WKL c
76 WKWN n
77 WRK c
78 YOEP x
79 ANC c
80 ANC1P x
81 ANC2P x
82 DECADE x
83 DIS x
84 DRIVESP x
85 ESP x
86 ESR x
87 FOD1P x
88 FOD2P x

89 HICOV x
90 HISP x
91 INDP x
92 JWAP x
93 JWDP x
94 LANP x
95 MIGPUMA x
96 MIGSP x
97 MSP c
98 NAICSP x
99 NATIVITY x
100 NOP x
101 OC c
102 OCCP x
103 PAOC x
104 PERNP n
105 PINCP x
106 POBP x
107 POVPIP n
108 POWPUMA x
109 POWSP x
110 PRIVCOV x
111 PUBCOV x
112 QTRBIR x
113 RAC1P x
114 RAC2P x
115 RAC3P x
116 RACAIAN x
117 RACASN x
118 RACBLK x
119 RACNH x
120 RACNUM x
121 RACPI x
122 RACSOR x
123 RACWHT x
124 RC c
125 SCIENGP x
126 SCIENGRLP x
127 SFN x
128 SFR x
129 SOCP x
130 VPS x
131 WAOB x
132 FAGEP x
133 FANCP x
134 FCITP x

135 FCITWP x
136 FCOWP x
137 FDDRSP x
138 FDEARP x
139 FDEYEP x
140 FDISP x
141 FDOUTP x
142 FDPHYP x
143 FDRATP x
144 FDRATXP x
145 FDREMP x
146 FENGP x
147 FESRP x
148 FFERP x
149 FFODP x
150 FGCLP x
151 FGCMP x
152 FGCRP x
153 FHICOVP x
154 FHIMRKSP x
155 FHINS1P x
156 FHINS2P x
157 FHINS3C x
158 FHINS3P x
159 FHINS4C x
160 FHINS4P x
161 FHINS5C x
162 FHINS5P x
163 FHINS6P x
164 FHINS7P x
165 FHISP x
166 FINDP x
167 FINTP d
168 FJWDP x
169 FJWMNP x
170 FJWRIP x
171 FJWTRNSP x
172 FLANP x
173 FLANXP x
174 FMARP x
175 FMARHDP x
176 FMARHMP x
177 FMARHTP x
178 FMARHWP x
179 FMARHYP x
180 FMIGP x

181 FMIGSP x
182 FMILPP x
183 FMILSP x
184 FOCCP x
185 FOIP x
186 FPAP x
187 FPERNP x
188 FPINCP x
189 FPOBP x
190 FPOWSP x
191 FPRIVCOVP x
192 FPUBCOVP x
193 FRACP x
194 FRELSHIPP x
195 FRETP x
196 FSCHGP x
197 FSCHLP x
198 FSCHP x
199 FSEMP x
200 FSEXP x
201 FSSIP x
202 FSSP x
203 FWAGP x
204 FWKHP x
205 FWKLP x
206 FWKWNP x
207 FWRKP x
208 FYOEP x
209 PWGTP1 x
210 PWGTP2 x
211 PWGTP3 x
212 PWGTP4 x
213 PWGTP5 x
214 PWGTP6 x
215 PWGTP7 x
216 PWGTP8 x
217 PWGTP9 x
218 PWGTP10 x
219 PWGTP11 x
220 PWGTP12 x
221 PWGTP13 x
222 PWGTP14 x
223 PWGTP15 x
224 PWGTP16 x
225 PWGTP17 x
226 PWGTP18 x

227 PWGTP19 x
228 PWGTP20 x
229 PWGTP21 x
230 PWGTP22 x
231 PWGTP23 x
232 PWGTP24 x
233 PWGTP25 x
234 PWGTP26 x
235 PWGTP27 x
236 PWGTP28 x
237 PWGTP29 x
238 PWGTP30 x
239 PWGTP31 x
240 PWGTP32 x
241 PWGTP33 x
242 PWGTP34 x
243 PWGTP35 x
244 PWGTP36 x
245 PWGTP37 x
246 PWGTP38 x
247 PWGTP39 x
248 PWGTP40 x
249 PWGTP41 x
250 PWGTP42 x
251 PWGTP43 x
252 PWGTP44 x
253 PWGTP45 x
254 PWGTP46 x
255 PWGTP47 x
256 PWGTP48 x
257 PWGTP49 x
258 PWGTP50 x
259 PWGTP51 x
260 PWGTP52 x
261 PWGTP53 x
262 PWGTP54 x
263 PWGTP55 x
264 PWGTP56 x
265 PWGTP57 x
266 PWGTP58 x
267 PWGTP59 x
268 PWGTP60 x
269 PWGTP61 x
270 PWGTP62 x
271 PWGTP63 x
272 PWGTP64 x

273 PWGTP65 x
274 PWGTP66 x
275 PWGTP67 x
276 PWGTP68 x
277 PWGTP69 x
278 PWGTP70 x
279 PWGTP71 x
280 PWGTP72 x
281 PWGTP73 x
282 PWGTP74 x
283 PWGTP75 x
284 PWGTP76 x
285 PWGTP77 x
286 PWGTP78 x
287 PWGTP79 x
288 PWGTP80 x

Description file for regression tree and forest after excluding unimportant variables

```
"data.txt"
 "NA"
 2
 1 RT x
 2 SERIALNO x
 3 DIVISION x
 4 SPORDER x
 5 PUMA x
 6 REGION x
 7 ST x
 8 ADJINC x
 9 PWGTP w
 10 AGEP n
 11 CIT x
 12 CITWP x
 13 COW c
 14 DDRS x
 15 DEAR x
 16 DEYE x
 17 DOUT x
 18 DPHY x
 19 DRAT x
 20 DRATX x
 21 DREM x
 22 ENG x
 23 FER c
 24 GCL c
 25 GCM x
 26 GCR x
 27 HIMRKS c
 28 HINS1 c
 29 HINS2 c
 30 HINS3 c
 31 HINS4 x
 32 HINS5 x
 33 HINS6 x
 34 HINS7 x
 35 INTP d
 36 JWMNP n
 37 JWRIP x
 38 JWTRNS c
 39 LANX x
 40 MAR c
 41 MARHD c
 42 MARHM c
```

43 MARHT c
44 MARHW c
45 MARHYP n
46 MIG x
47 MIL x
48 MLPA x
49 MLPB x
50 MLPCD x
51 MLPE x
52 MLPFG x
53 MLPH x
54 MLPI x
55 MLPJ x
56 MLPK x
57 NWAB c
58 NWAV x
59 NWLA c
60 NWLK c
61 NWRE x
62 OIP x
63 PAP x
64 RELSHIPP c
65 RETP n
66 SCH c
67 SCHG x
68 SCHL c
69 SEMP x
70 SEX c
71 SSIP x
72 SSP n
73 WAGP n
74 WKHP n
75 WKL c
76 WKWN n
77 WRK c
78 YOEP x
79 ANC x
80 ANC1P x
81 ANC2P x
82 DECADE x
83 DIS x
84 DRIVESP x
85 ESP x
86 ESR c
87 FOD1P x
88 FOD2P x

89 HICOV x
90 HISP x
91 INDP x
92 JWAP n
93 JWDP n
94 LANP x
95 MIGPUMA x
96 MIGSP x
97 MSP c
98 NAICSP x
99 NATIVITY x
100 NOP x
101 OC c
102 OCCP x
103 PAOC x
104 PERNP n
105 PINCP n
106 POBP x
107 POVPIP n
108 POWPUMA x
109 POWSP x
110 PRIVCOV x
111 PUBCOV c
112 QTRBIR x
113 RAC1P x
114 RAC2P x
115 RAC3P x
116 RACAIAN x
117 RACASN x
118 RACBLK x
119 RACNH c
120 RACNUM x
121 RACPI x
122 RACSOR x
123 RACWHT x
124 RC c
125 SCIENGP x
126 SCIENGRLP x
127 SFN x
128 SFR x
129 SOCP x
130 VPS x
131 WAOB x
132 FAGEP x
133 FANCP x
134 FCITP x

135 FCITWP x
136 FCOWP x
137 FDDRSP x
138 FDEARP x
139 FDEYEP x
140 FDISP x
141 FDOUTP x
142 FDPHYP x
143 FDRATP x
144 FDRATXP x
145 FDREMP x
146 FENGP x
147 FESRP x
148 FFERP x
149 FFODP x
150 FGCLP x
151 FGCMP x
152 FGCRP x
153 FHICOVP x
154 FHIMRKSP x
155 FHINS1P x
156 FHINS2P x
157 FHINS3C x
158 FHINS3P x
159 FHINS4C x
160 FHINS4P x
161 FHINS5C x
162 FHINS5P x
163 FHINS6P x
164 FHINS7P x
165 FHISP x
166 FINDP x
167 FINTP c
168 FJWDP x
169 FJWMNP x
170 FJWRIP x
171 FJWTRNSP x
172 FLANP x
173 FLANXP x
174 FMARP x
175 FMARHDP x
176 FMARHMP x
177 FMARHTP x
178 FMARHWP x
179 FMARHYP x
180 FMIGP x

181 FMIGSP x
182 FMILPP x
183 FMILSP x
184 FOCCP x
185 FOIP x
186 FPAP x
187 FPERNP x
188 FPINCP x
189 FPOBP x
190 FPOWSP x
191 FPRIVCOVP x
192 FPUBCOVP x
193 FRACP x
194 FRELSHIPP x
195 FRETP x
196 FSCHGP x
197 FSCHLP x
198 FSCHP x
199 FSEMP x
200 FSEXP x
201 FSSIP x
202 FSSP x
203 FWAGP x
204 FWKHP x
205 FWKLP x
206 FWKWNP x
207 FWRKP x
208 FYOEP x
209 PWGTP1 x
210 PWGTP2 x
211 PWGTP3 x
212 PWGTP4 x
213 PWGTP5 x
214 PWGTP6 x
215 PWGTP7 x
216 PWGTP8 x
217 PWGTP9 x
218 PWGTP10 x
219 PWGTP11 x
220 PWGTP12 x
221 PWGTP13 x
222 PWGTP14 x
223 PWGTP15 x
224 PWGTP16 x
225 PWGTP17 x
226 PWGTP18 x

227 PWGTP19 x
228 PWGTP20 x
229 PWGTP21 x
230 PWGTP22 x
231 PWGTP23 x
232 PWGTP24 x
233 PWGTP25 x
234 PWGTP26 x
235 PWGTP27 x
236 PWGTP28 x
237 PWGTP29 x
238 PWGTP30 x
239 PWGTP31 x
240 PWGTP32 x
241 PWGTP33 x
242 PWGTP34 x
243 PWGTP35 x
244 PWGTP36 x
245 PWGTP37 x
246 PWGTP38 x
247 PWGTP39 x
248 PWGTP40 x
249 PWGTP41 x
250 PWGTP42 x
251 PWGTP43 x
252 PWGTP44 x
253 PWGTP45 x
254 PWGTP46 x
255 PWGTP47 x
256 PWGTP48 x
257 PWGTP49 x
258 PWGTP50 x
259 PWGTP51 x
260 PWGTP52 x
261 PWGTP53 x
262 PWGTP54 x
263 PWGTP55 x
264 PWGTP56 x
265 PWGTP57 x
266 PWGTP58 x
267 PWGTP59 x
268 PWGTP60 x
269 PWGTP61 x
270 PWGTP62 x
271 PWGTP63 x
272 PWGTP64 x

273 PWGTP65 x
274 PWGTP66 x
275 PWGTP67 x
276 PWGTP68 x
277 PWGTP69 x
278 PWGTP70 x
279 PWGTP71 x
280 PWGTP72 x
281 PWGTP73 x
282 PWGTP74 x
283 PWGTP75 x
284 PWGTP76 x
285 PWGTP77 x
286 PWGTP78 x
287 PWGTP79 x
288 PWGTP80 x

Text of RPART regression tree

node), split, n, deviance, yval
* denotes terminal node

1) root 39656 1.073711e+15 2007.620
2) PINCP< 253500 39063 1.181941e+14 996.044 *
3) PINCP>=253500 593 7.071150e+14 64909.590
6) PERNP>=116500 428 1.731330e+14 19369.620
12) PINCP< 706450 409 8.285635e+13 11660.140
24) PERNP>=232500 387 2.849638e+13 6276.037 *
25) PERNP< 232500 22 3.729885e+13 78798.520
50) PINCP< 375500 11 1.539628e+12 20389.670 *
51) PINCP>=375500 11 5.957898e+11 250608.500 *
13) PINCP>=706450 19 2.166468e+13 200220.700 *
7) PERNP< 116500 165 3.261137e+13 243093.700
14) RETP>=111000 11 1.506378e+13 115166.000 *
15) RETP< 111000 154 0.000000e+00 254000.000 *

Text of RPART classification tree

```
n= 46693

node), split, n, loss, yval, (yprob)
      * denotes terminal node

  1) root 46693 7037 0 (0.8492922 0.1507078)
    2) ANC=1,2,3 39570 3350 0 (0.9153399 0.0846601) *
    3) ANC=4 7123 3436 1 (0.4823810 0.5176190)
      6) NWLK=1,2 1460  307 0 (0.7897260 0.2102740) *
      7) NWLK=3 5663 2283 1 (0.4031432 0.5968568)
       14) ESR=1,2,4 3910 1947 1 (0.4979540 0.5020460)
         28) RELSHIPP=20,21,22,23,24,27,29,35 2920 1328 0 (0.5452055 0.4547945)
           56) RACWHT=1 2741 1194 0 (0.5643926 0.4356074)
            112) AGEP< 49.5 1751  661 0 (0.6225014 0.3774986) *
            113) AGEP>=49.5 990  457 1 (0.4616162 0.5383838) *
           57) RACWHT=0 179   45 1 (0.2513966 0.7486034) *
         29) RELSHIPP=25,26,28,30,31,32,33,34,36,38 990  355 1 (0.3585859 0.6414141) *
       15) ESR=3,6 1753  336 1 (0.1916714 0.8083286) *
```

Text of CTREE regression tree

Conditional inference tree with 266 terminal nodes

Response: INTP
Inputs: CIT, COW, DDRS, DEAR, DEYE, DOUT, DPHY, DRAT, DRATX, DREM, ENG, FER, GCL, GCM, GCR, HIMRKS, HINS1, HINS2, HINS3, HINS4, HINS5, HINS6, HINS7, JWTRNS, LANX, MAR, MARHD, MARHM, MARHT, MARHW, MIG, MIL, MLPA, MLPB, MLPCD, MLPE, MLPFG, MLPH, MLPI, MLPJ, MLPK, NWAB, NWAV, NWLA, NWLK, NWRE, RELSHIPP, SCH, SCHG, SCHL, SEX, WKL, WRK, ANC, DIS, ESP, ESR, HICOV, HISP, MIGPUMA, MSP, NATIVITY, NOP, OC, OCCP, PAOC, POWPUMA, PRIVCOV, PUBCOV, QTRBIR, RAC1P, RACAIAN, RACASN, RACBLK, RACNH, RACPI, RACSOR, RACWHT, RC, SFN, SFR, VPS, WAOB, FINTP, PWGTP, AGEP, CITWP, JWMNP, JWRIP, MARHYP, OIP, PAP, RETP, SEMP, SSIP, SSP, WAGP, WKHP, WKWN, YOEP, JWAP, JWDP, PERNP, PINCP, POVPIP, RACNUM
Number of observations: 39656

1) PINCP <= 253000; criterion = 1, statistic = 5901.35
  2) PINCP <= 80070; criterion = 1, statistic = 1686.192
    3) AGEP <= 61; criterion = 1, statistic = 1104.316
      4) MIGPUMA == {00100, 00190, 00500, 06800}; criterion = 1, statistic = 1034.042
        5) WAOB == {3, 4, 5, 6, 7}; criterion = 1, statistic = 96.313
          6)* weights = 9
        5) WAOB == {1}
          7) SEMP <= 15000; criterion = 1, statistic = 26.728
            8) WKL == {2, 3}; criterion = 1, statistic = 24.63
              9)* weights = 18
            8) WKL == {1}
              10)* weights = 148
          7) SEMP > 15000
            11)* weights = 7
      4) MIGPUMA == {00001, 00002, 00104, 00105, 00200, 00300, 00390, 00400, 00490, 00590, 00600, 00690, 00700, 00702, 00790, 00800, 00900, 01000, 01001, 01100, 01200, 01290, 01300, 01301, 01325, 01327, 01400, 01401, 01490, 01500, 01501, 01600, 01601, 01690, 01700, 01800, 01900, 02000, 02100, 02101, 02102, 02200, 02300, 02400, 02500, 02600, 02700, 02890, 02900, 03000, 03100, 03200, 03300, 03400, 03500, 03600, 03700, 03800, 04000, 04007, 04100, 04500, 04600, 04690, 05690, 05700, 05900, 06900, 07100, 07300, 07900, 08100, 08500, 08690, 09500, 09900, 10000, 10100, 10300, 10700, 11100, 11600, 20000, 30000, 35000, 40100, 51085, 55100, 55102, 59300, 70100}
        12) AGEP <= 48; criterion = 1, statistic = 239.262
          13) PINCP <= 38700; criterion = 1, statistic = 72.338
            14) AGEP <= 33; criterion = 1, statistic = 64.775
              15)* weights = 7029
            14) AGEP > 33
              16) NWAV == {3}; criterion = 1, statistic = 36.787
                17)* weights = 60
              16) NWAV == {1, 2, 5}

18) SCH == {2, 3}; criterion = 0.994, statistic = 42.782
19)* weights = 137
18) SCH == {1}
20) POWPUMA == {00400, 01900, 02400}; criterion = 0.961, statistic = 63.059
21)* weights = 164
20) POWPUMA == {00001, 00100, 00190, 00200, 00300, 00390, 00500, 00590, 00600, 00700, 00800, 00900, 01000, 01100, 01200, 01300, 01400, 01500, 01690, 01800, 02000, 02100, 02200, 02300, 02500, 02600, 03200, 55102}
22) MIGPUMA == {01300, 02200, 03700}; criterion = 1, statistic = 109.732
23)* weights = 32
22) MIGPUMA == {00001, 00200, 00390, 00400, 00590, 00600, 00700, 00800, 00900, 01000, 01100, 01200, 01400, 01490, 01500, 01600, 01690, 01800, 01900, 02000, 02100, 02102, 02300, 02400, 02500, 02600, 02700, 02900, 03300, 03500, 04100, 05690, 05700, 07300, 09900, 10700, 11600, 40100, 55100, 70100}
24) WKL == {3}; criterion = 0.997, statistic = 28.233
25) PINCP <= 8500; criterion = 0.993, statistic = 15.246
26)* weights = 267
25) PINCP > 8500
27) PUBCOV == {2}; criterion = 1, statistic = 21.608
28)* weights = 11
27) PUBCOV == {1}
29)* weights = 75
24) WKL == {1, 2}
30)* weights = 2287
13) PINCP > 38700
31) PERNP <= 28000; criterion = 1, statistic = 378.054
32) SCHL == {14, 21, 22}; criterion = 0.972, statistic = 27.002
33)* weights = 14
32) SCHL == {16, 17, 18, 19, 20}
34)* weights = 49
31) PERNP > 28000
35) COW == {6, 7}; criterion = 1, statistic = 62.605
36) HISP == {02, 04, 16, 17}; criterion = 1, statistic = 42.461
37)* weights = 8
36) HISP == {01}
38) NWAV == {1, 3}; criterion = 1, statistic = 34.505
39)* weights = 11
38) NWAV == {5}
40)* weights = 361
35) COW == {1, 2, 3, 4, 5, 8}
41) WAGP <= 38000; criterion = 0.997, statistic = 50.188
42) WAOB == {4, 5, 6}; criterion = 0.975, statistic = 18.754
43)* weights = 7
42) WAOB == {1}
44) MAR == {3, 4}; criterion = 0.998, statistic = 27.213
45)* weights = 18

```
              44) MAR == {1, 2, 5}
                46)*  weights = 82
            41) WAGP > 38000
              47) PINCP <= 60800; criterion = 0.998, statistic = 50.839
                48) MIG == {1, 2}; criterion = 1, statistic = 41.586
                  49) HISP == {04, 05, 06, 07, 08, 09, 11, 13, 15, 16, 17, 19, 20, 21, 23, 24};
criterion = 0.984, statistic = 47.777
                    50)*  weights = 29
                  49) HISP == {01, 02, 03}
                    51) HIMRKS == {2}; criterion = 0.958, statistic = 35.256
                      52)*  weights = 121
                    51) HIMRKS == {0, 1}
                      53)*  weights = 2639
                48) MIG == {3}
                  54) SCH == {2, 3}; criterion = 1, statistic = 79.057
                    55)*  weights = 28
                  54) SCH == {1}
                    56) RAC1P == {3, 6}; criterion = 1, statistic = 184.549
                      57)*  weights = 15
                    56) RAC1P == {1, 2, 8, 9}
                      58) MIL == {1, 3, 4}; criterion = 0.95, statistic = 17.401
                        59)*  weights = 415
                      58) MIL == {2}
                        60)*  weights = 10
              47) PINCP > 60800
                61) WAGP <= 51000; criterion = 1, statistic = 99.129
                  62)*  weights = 33
                61) WAGP > 51000
                  63) WAGP <= 58000; criterion = 1, statistic = 54.201
                    64) JWAP <= 93; criterion = 0.965, statistic = 21.265
                      65)*  weights = 22
                    64) JWAP > 93
                      66)*  weights = 8
                  63) WAGP > 58000
                    67) OCCP == {1, 3, 4, 5, 6, 9, 18, 19, 21}; criterion = 0.993, statistic = 57.984
                      68)*  weights = 690
                    67) OCCP == {2, 7, 8, 10, 11, 12, 13, 14, 15, 16, 17, 20, 22, 23, 24}
                      69) NWRE == {1, 2}; criterion = 0.975, statistic = 28.737
                        70)*  weights = 16
                      69) NWRE == {3}
                        71)*  weights = 894
        12) AGEP > 48
          72) POWPUMA == {01000, 01690, 01900, 02000, 02300, 02600, 03400}; criterion = 1,
statistic = 167.902
            73) POWPUMA == {01000, 03400}; criterion = 1, statistic = 62.679
              74) PWGTP <= 238; criterion = 0.988, statistic = 22.799
```

```
          75)*  weights = 65
        74) PWGTP > 238
          76)*  weights = 7
      73) POWPUMA == {01690, 01900, 02000, 02300, 02600}
        77) PINCP <= 60400; criterion = 0.974, statistic = 25.141
          78) JWTRNS == {09, 10, 12}; criterion = 1, statistic = 36.404
            79) COW == {1, 3}; criterion = 1, statistic = 31.944
              80)*  weights = 14
            79) COW == {2, 4, 6, 7}
              81)*  weights = 25
          78) JWTRNS == {01, 02, 11}
            82) HINS3 == {1}; criterion = 0.984, statistic = 26.891
              83)*  weights = 9
            82) HINS3 == {2}
              84) COW == {7}; criterion = 0.996, statistic = 32.078
                85)*  weights = 35
              84) COW == {1, 2, 3, 4, 5, 6, 8}
                86)*  weights = 688
        77) PINCP > 60400
          87) PERNP <= 56000; criterion = 1, statistic = 81.587
            88)*  weights = 20
          87) PERNP > 56000
            89) CIT == {2, 3, 4, 5}; criterion = 0.964, statistic = 20.282
              90)*  weights = 8
            89) CIT == {1}
              91)*  weights = 167
    72) POWPUMA == {00100, 00104, 00200, 00300, 00390, 00400, 00500, 00590, 00600,
00700, 00701, 00800, 00900, 01100, 01200, 01300, 01400, 01490, 01500, 01700, 01800, 02100,
02200, 02400, 02500, 02800, 03001, 03200, 03800, 55100, 55102}
      92) HIMRKS == {1, 2}; criterion = 1, statistic = 107.671
        93) PINCP <= 46800; criterion = 1, statistic = 34.091
          94) JWRIP <= 1; criterion = 0.999, statistic = 33.515
            95) WAGP <= 2300; criterion = 0.964, statistic = 32.867
              96) PINCP <= 23600; criterion = 1, statistic = 29.983
                97) MSP == {2, 3, 5}; criterion = 1, statistic = 48.456
                  98)*  weights = 8
                97) MSP == {1, 4, 6}
                  99)*  weights = 226
              96) PINCP > 23600
                100)*  weights = 63
            95) WAGP > 2300
              101) JWTRNS == {02, 10}; criterion = 0.999, statistic = 43.271
                102)*  weights = 14
              101) JWTRNS == {01, 09, 11, 12}
                103) ESR == {6}; criterion = 1, statistic = 45.455
                  104)*  weights = 19
```

```
          103) ESR == {1, 2, 3}
            105)*  weights = 274
        94) JWRIP > 1
          106)*  weights = 18
      93) PINCP > 46800
        107) PERNP <= 39700; criterion = 1, statistic = 45.738
          108) RETP <= 16800; criterion = 0.958, statistic = 11.888
            109) OIP <= 750; criterion = 0.995, statistic = 15.625
              110)*  weights = 15
            109) OIP > 750
              111)*  weights = 8
          108) RETP > 16800
            112)*  weights = 19
        107) PERNP > 39700
          113)*  weights = 166
    92) HIMRKS == {0}
      114) WRK == {2}; criterion = 1, statistic = 31.53
        115) PINCP <= 22800; criterion = 1, statistic = 59.836
          116) PINCP <= 2900; criterion = 1, statistic = 37.751
            117) SEMP <= 0; criterion = 1, statistic = 128.799
              118) PERNP <= 450; criterion = 1, statistic = 145.119
                119) PERNP <= 0; criterion = 1, statistic = 145.881
                  120) WKL == {1, 3}; criterion = 1, statistic = 143.494
                    121)*  weights = 326
                  120) WKL == {2}
                    122) RELSHIPP == {20, 21, 22, 25, 33, 36}; criterion = 1, statistic = 59.949
                      123)*  weights = 175
                    122) RELSHIPP == {23, 29, 34, 38}
                      124)*  weights = 9
                119) PERNP > 0
                  125)*  weights = 10
              118) PERNP > 450
                126)*  weights = 17
            117) SEMP > 0
              127)*  weights = 9
          116) PINCP > 2900
            128) DIS == {2}; criterion = 1, statistic = 36.426
              129)*  weights = 249
            128) DIS == {1}
              130) NWAV == {2}; criterion = 0.969, statistic = 18.204
                131)*  weights = 7
              130) NWAV == {1, 3, 5}
                132) HINS6 == {1}; criterion = 0.997, statistic = 17.134
                  133)*  weights = 8
                132) HINS6 == {2}
                  134)*  weights = 283
```

```
          115) PINCP > 22800
            135)*  weights = 332
          114) WRK == {1}
            136) PINCP <= 62400; criterion = 1, statistic = 27.807
              137) JWTRNS == {08, 10}; criterion = 0.988, statistic = 31.71
                138) MIL == {2, 3}; criterion = 0.982, statistic = 16.828
                  139)*  weights = 9
                138) MIL == {4}
                  140)*  weights = 48
              137) JWTRNS == {01, 02, 05, 07, 09, 11, 12}
                141)*  weights = 3245
            136) PINCP > 62400
              142) PERNP <= 45000; criterion = 1, statistic = 166.058
                143)*  weights = 23
              142) PERNP > 45000
                144) PERNP <= 62000; criterion = 1, statistic = 70.17
                  145)*  weights = 53
                144) PERNP > 62000
                  146)*  weights = 751
3) AGEP > 61
  147) PINCP <= 38420; criterion = 1, statistic = 868.424
    148) PINCP <= 21880; criterion = 1, statistic = 312.715
      149) PINCP <= 14470; criterion = 1, statistic = 87.755
        150) SCHL == {18, 21, 22, 23, 24}; criterion = 1, statistic = 116.754
          151) SSP <= 6000; criterion = 1, statistic = 31.543
            152) PINCP <= 6600; criterion = 1, statistic = 49.342
              153)*  weights = 237
            152) PINCP > 6600
              154) PERNP <= 1700; criterion = 0.995, statistic = 21.067
                155)*  weights = 69
              154) PERNP > 1700
                156)*  weights = 42
          151) SSP > 6000
            157) CIT == {3, 4}; criterion = 0.963, statistic = 19.576
              158)*  weights = 7
            157) CIT == {1, 5}
              159)*  weights = 340
        150) SCHL == {01, 03, 07, 08, 09, 10, 11, 12, 13, 14, 15, 16, 17, 19, 20}
          160) SSP <= 7300; criterion = 0.996, statistic = 37.339
            161) PINCP <= 9760; criterion = 1, statistic = 42.695
              162) OCCP == {2, 10}; criterion = 1, statistic = 71.233
                163)*  weights = 8
              162) OCCP == {1, 3, 4, 5, 6, 9, 11, 12, 14, 15, 16, 17, 18, 19, 20, 22, 23, 24, 9920}
                164)*  weights = 909
            161) PINCP > 9760
              165)*  weights = 205
```

```
          160) SSP > 7300
            166) PINCP <= 10100; criterion = 0.968, statistic = 26.249
              167) RELSHIPP == {21, 22, 23, 25, 29, 31, 33, 34, 36, 37, 38}; criterion = 1, statistic
= 62.664
                168)*  weights = 268
              167) RELSHIPP == {20, 28}
                169)*  weights = 136
            166) PINCP > 10100
              170) SSP <= 10000; criterion = 1, statistic = 35.589
                171) RETP <= 0; criterion = 0.993, statistic = 15.421
                  172)*  weights = 72
                171) RETP > 0
                  173)*  weights = 79
              170) SSP > 10000
                174) MIGPUMA == {00390, 01100, 01690, 02000, 02200, 02600, 06500, 10500,
10700}; criterion = 0.995, statistic = 57.02
                  175)*  weights = 13
                174) MIGPUMA == {00100, 00200, 00600, 00700, 00800, 00900, 01200, 01300,
01400, 01500, 01800, 02100, 02300, 02400}
                  176)*  weights = 660
        149) PINCP > 14470
          177) SSP <= 12900; criterion = 1, statistic = 71.651
            178) PERNP <= 3200; criterion = 1, statistic = 38.747
              179) RETP <= 1600; criterion = 1, statistic = 65.894
                180) OIP <= 1700; criterion = 1, statistic = 25.048
                  181) SSIP <= 0; criterion = 1, statistic = 39.901
                    182) SSP <= 7900; criterion = 1, statistic = 48.762
                      183)*  weights = 22
                    182) SSP > 7900
                      184) PINCP <= 18600; criterion = 1, statistic = 34.089
                        185) WAGP <= 1000; criterion = 0.991, statistic = 14.808
                          186)*  weights = 37
                        185) WAGP > 1000
                          187)*  weights = 8
                      184) PINCP > 18600
                        188)*  weights = 18
                  181) SSIP > 0
                    189)*  weights = 37
                180) OIP > 1700
                  190)*  weights = 49
              179) RETP > 1600
                191) SSP <= 6000; criterion = 0.999, statistic = 35.681
                  192) RETP <= 9000; criterion = 1, statistic = 22.922
                    193)*  weights = 15
                  192) RETP > 9000
                    194)*  weights = 61
```

191) SSP > 6000
195) RETP <= 5000; criterion = 1, statistic = 23.987
196) PINCP <= 18300; criterion = 1, statistic = 34.09
197) SSP <= 10000; criterion = 0.999, statistic = 19.536
198)* weights = 9
197) SSP > 10000
199)* weights = 51
196) PINCP > 18300
200)* weights = 13
195) RETP > 5000
201)* weights = 195
178) PERNP > 3200
202) POWPUMA == {02100, 09900}; criterion = 1, statistic = 89.257
203)* weights = 13
202) POWPUMA == {00100, 00200, 00390, 00400, 00500, 00600, 00700, 00800, 00900, 01000, 01100, 01200, 01300, 01400, 01500, 01690, 01800, 01900, 02000, 02200, 02300, 02400, 02500, 02600}
204) JWTRNS == {02, 08, 09, 12}; criterion = 1, statistic = 159.191
205)* weights = 9
204) JWTRNS == {01, 10, 11}
206) OCCP == {5, 13, 24}; criterion = 1, statistic = 112.168
207)* weights = 26
206) OCCP == {1, 2, 3, 4, 7, 9, 10, 11, 12, 14, 15, 16, 17, 18, 19, 20, 22, 23}
208) SCHL == {01, 11, 14, 15, 16, 17, 19, 20, 21, 22, 23, 24}; criterion = 0.998, statistic = 42.443
209) SCHL == {15, 21}; criterion = 1, statistic = 52.045
210)* weights = 33
209) SCHL == {01, 11, 14, 16, 17, 19, 20, 22, 23, 24}
211) NWAV == {1, 3}; criterion = 0.978, statistic = 16.463
212)* weights = 11
211) NWAV == {5}
213) JWTRNS == {01, 11}; criterion = 0.959, statistic = 15.183
214) OCCP == {3, 14}; criterion = 0.996, statistic = 47.725
215)* weights = 18
214) OCCP == {1, 4, 7, 9, 10, 11, 12, 15, 16, 17, 18, 19, 20, 22, 23}
216)* weights = 116
213) JWTRNS == {10}
217)* weights = 7
208) SCHL == {18}
218)* weights = 31
177) SSP > 12900
219) SSP <= 14400; criterion = 1, statistic = 50.875
220)* weights = 163
219) SSP > 14400
221)* weights = 719
148) PINCP > 21880

```
222) RETP <= 5000; criterion = 1, statistic = 76.748
  223) WKL == {2, 3}; criterion = 1, statistic = 138.678
    224) SSP <= 16800; criterion = 1, statistic = 163.95
      225) OIP <= 1600; criterion = 1, statistic = 77.59
        226) SSIP <= 1500; criterion = 1, statistic = 69.28
          227) SSP <= 9800; criterion = 1, statistic = 63.062
            228) PINCP <= 31000; criterion = 1, statistic = 22.992
              229) SSP <= 6400; criterion = 0.969, statistic = 12.329
                230)*  weights = 16
              229) SSP > 6400
                231)*  weights = 12
            228) PINCP > 31000
              232)*  weights = 16
          227) SSP > 9800
            233) PINCP <= 30300; criterion = 1, statistic = 56.354
              234) SSP <= 12200; criterion = 0.993, statistic = 18.675
                235)*  weights = 17
              234) SSP > 12200
                236)*  weights = 53
            233) PINCP > 30300
              237)*  weights = 27
        226) SSIP > 1500
          238)*  weights = 35
      225) OIP > 1600
        239)*  weights = 94
    224) SSP > 16800
      240) PINCP <= 30200; criterion = 1, statistic = 35.893
        241) SSP <= 19100; criterion = 1, statistic = 60.754
          242) RETP <= 2500; criterion = 0.995, statistic = 15.881
            243) OIP <= 3100; criterion = 0.963, statistic = 12.019
              244)*  weights = 24
            243) OIP > 3100
              245)*  weights = 7
          242) RETP > 2500
            246)*  weights = 19
        241) SSP > 19100
          247) SSP <= 21600; criterion = 1, statistic = 34.308
            248) RETP <= 0; criterion = 0.999, statistic = 19.173
              249)*  weights = 28
            248) RETP > 0
              250)*  weights = 40
          247) SSP > 21600
            251) OCCP == {3, 10, 20}; criterion = 1, statistic = 55.433
              252)*  weights = 12
            251) OCCP == {1, 2, 4, 5, 9, 11, 12, 13, 15, 17, 18, 19, 22, 23, 24}
              253) COW == {2, 3, 5}; criterion = 1, statistic = 39.272
```

```
            254)*  weights = 10
          253) COW == {1, 6, 7, 8}
            255) PINCP <= 26400; criterion = 0.959, statistic = 17.652
              256)*  weights = 146
            255) PINCP > 26400
              257)*  weights = 65
      240) PINCP > 30200
        258) SSP <= 29000; criterion = 1, statistic = 21.786
          259) OIP <= 6900; criterion = 0.993, statistic = 15.087
            260)*  weights = 44
          259) OIP > 6900
            261)*  weights = 15
        258) SSP > 29000
          262)*  weights = 39
    223) WKL == {1}
      263) PERNP <= 1700; criterion = 1, statistic = 77.279
        264) SSP <= 10900; criterion = 1, statistic = 22.086
          265)*  weights = 7
        264) SSP > 10900
          266)*  weights = 20
      263) PERNP > 1700
        267) WAGP <= 6000; criterion = 1, statistic = 66.445
          268)*  weights = 161
        267) WAGP > 6000
          269) SCHL == {21, 22, 23, 24}; criterion = 0.991, statistic = 49.775
            270)*  weights = 105
          269) SCHL == {01, 05, 06, 07, 08, 09, 12, 13, 14, 15, 16, 17, 18, 19, 20}
            271)*  weights = 522
  222) RETP > 5000
    272) SSP <= 11400; criterion = 1, statistic = 63.523
      273) RETP <= 12000; criterion = 0.999, statistic = 40.265
        274) WKL == {2, 3}; criterion = 0.984, statistic = 27.517
          275) PINCP <= 28000; criterion = 0.997, statistic = 23.947
            276)*  weights = 26
          275) PINCP > 28000
            277)*  weights = 12
        274) WKL == {1}
          278)*  weights = 38
      273) RETP > 12000
        279)*  weights = 221
    272) SSP > 11400
      280) RETP <= 8800; criterion = 0.993, statistic = 25.642
        281) AGEP <= 73; criterion = 0.999, statistic = 19.347
          282)*  weights = 146
        281) AGEP > 73
          283) PINCP <= 27000; criterion = 0.99, statistic = 22.229
```

```
          284)*  weights = 55
        283) PINCP > 27000
          285) SSP <= 21100; criterion = 0.98, statistic = 21.416
            286)*  weights = 38
          285) SSP > 21100
            287)*  weights = 21
      280) RETP > 8800
        288) PINCP <= 35920; criterion = 0.995, statistic = 19.749
          289) WAOB == {3, 4, 7}; criterion = 1, statistic = 36.043
            290)*  weights = 7
          289) WAOB == {1, 5}
            291)*  weights = 561
        288) PINCP > 35920
          292) RETP <= 10800; criterion = 0.956, statistic = 28.104
            293)*  weights = 18
          292) RETP > 10800
            294)*  weights = 156
147) PINCP > 38420
  295) PERNP <= 11300; criterion = 1, statistic = 185.194
    296) RETP <= 10900; criterion = 1, statistic = 430.57
      297) OIP <= 6500; criterion = 1, statistic = 150.678
        298) SSP <= 19200; criterion = 1, statistic = 162.985
          299) PINCP <= 56000; criterion = 1, statistic = 102.804
            300) SSP <= 8000; criterion = 1, statistic = 60.098
              301) PINCP <= 46800; criterion = 0.999, statistic = 19.115
                302)*  weights = 14
              301) PINCP > 46800
                303)*  weights = 20
            300) SSP > 8000
              304) SSIP <= 0; criterion = 1, statistic = 34.187
                305) PINCP <= 44200; criterion = 1, statistic = 30.963
                  306) RETP <= 7200; criterion = 0.978, statistic = 13.04
                    307)*  weights = 32
                  306) RETP > 7200
                    308)*  weights = 7
                305) PINCP > 44200
                  309) RETP <= 2600; criterion = 1, statistic = 23.203
                    310) PINCP <= 48600; criterion = 0.973, statistic = 12.568
                      311)*  weights = 14
                    310) PINCP > 48600
                      312)*  weights = 15
                  309) RETP > 2600
                    313)*  weights = 18
              304) SSIP > 0
                314)*  weights = 7
          299) PINCP > 56000
```

```
      315) PINCP <= 63000; criterion = 0.997, statistic = 17.058
        316)*  weights = 24
      315) PINCP > 63000
        317)*  weights = 31
    298) SSP > 19200
      318) PINCP <= 57400; criterion = 1, statistic = 62.53
      319) SSP <= 27000; criterion = 1, statistic = 32.738
        320) PINCP <= 42900; criterion = 0.998, statistic = 17.488
          321) PERNP <= 930; criterion = 0.962, statistic = 11.945
            322)*  weights = 19
          321) PERNP > 930
            323)*  weights = 9
        320) PINCP > 42900
          324)*  weights = 27
      319) SSP > 27000
        325) PINCP <= 44300; criterion = 0.966, statistic = 12.287
          326)*  weights = 30
        325) PINCP > 44300
          327)*  weights = 22
    318) PINCP > 57400
      328)*  weights = 17
  297) OIP > 6500
    329)*  weights = 166
296) RETP > 10900
  330) PINCP <= 62900; criterion = 1, statistic = 124.128
    331) RETP <= 16500; criterion = 1, statistic = 71.713
      332) PINCP <= 47800; criterion = 1, statistic = 23.131
        333)*  weights = 86
      332) PINCP > 47800
        334) OIP <= 0; criterion = 0.965, statistic = 12.189
          335) SSP <= 27600; criterion = 0.998, statistic = 17.278
            336)*  weights = 31
          335) SSP > 27600
            337)*  weights = 11
        334) OIP > 0
          338)*  weights = 18
    331) RETP > 16500
      339) PINCP <= 53000; criterion = 1, statistic = 32.188
        340) SSP <= 16500; criterion = 0.999, statistic = 26.166
          341) RETP <= 22500; criterion = 1, statistic = 24.377
            342)*  weights = 31
          341) RETP > 22500
            343) NWLA == {1, 3}; criterion = 0.999, statistic = 25.705
              344)*  weights = 20
            343) NWLA == {2}
              345)*  weights = 238
```

```
          340) SSP > 16500
            346) RETP <= 17000; criterion = 0.984, statistic = 40.029
              347)*  weights = 13
            346) RETP > 17000
              348)*  weights = 435
          339) PINCP > 53000
            349)*  weights = 339
        330) PINCP > 62900
          350) RETP <= 24400; criterion = 1, statistic = 100.469
            351) OIP <= 15800; criterion = 1, statistic = 24.481
              352) SSP <= 23800; criterion = 1, statistic = 21.895
                353)*  weights = 25
              352) SSP > 23800
                354)*  weights = 10
            351) OIP > 15800
              355)*  weights = 11
          350) RETP > 24400
            356) RETP <= 37000; criterion = 1, statistic = 38.262
              357) SSP <= 16900; criterion = 1, statistic = 25.394
                358)*  weights = 18
              357) SSP > 16900
                359)*  weights = 56
            356) RETP > 37000
              360) SSP <= 18300; criterion = 0.977, statistic = 25.859
                361) RETP <= 51000; criterion = 0.998, statistic = 17.764
                  362)*  weights = 26
                361) RETP > 51000
                  363)*  weights = 68
              360) SSP > 18300
                364) COW == {3, 6, 7}; criterion = 0.989, statistic = 31.81
                  365)*  weights = 15
                364) COW == {1, 2, 4}
                  366)*  weights = 109
    295) PERNP > 11300
      367) PERNP <= 31600; criterion = 1, statistic = 50.313
        368) SSP <= 13500; criterion = 1, statistic = 33.47
          369) PINCP <= 68410; criterion = 1, statistic = 23.531
            370) RETP <= 10000; criterion = 0.991, statistic = 26.93
              371) PINCP <= 42800; criterion = 0.993, statistic = 21.54
                372)*  weights = 31
              371) PINCP > 42800
                373)*  weights = 31
            370) RETP > 10000
              374)*  weights = 47
          369) PINCP > 68410
            375)*  weights = 11
```

```
          368) SSP > 13500
            376) NWAV == {1, 3}; criterion = 0.999, statistic = 36.171
              377)*  weights = 17
            376) NWAV == {2, 5}
              378)*  weights = 205
          367) PERNP > 31600
            379) PERNP <= 43000; criterion = 0.996, statistic = 37.369
              380) ESR == {2, 3}; criterion = 1, statistic = 67.892
                381)*  weights = 9
              380) ESR == {1, 6}
                382) PINCP <= 61300; criterion = 1, statistic = 31.425
                  383) SCHL == {13, 14, 21, 23}; criterion = 1, statistic = 47.917
                    384)*  weights = 47
                  383) SCHL == {11, 12, 15, 16, 17, 18, 19, 20, 22}
                    385)*  weights = 209
                382) PINCP > 61300
                  386) HINS3 == {2}; criterion = 0.999, statistic = 19.716
                    387)*  weights = 7
                  386) HINS3 == {1}
                    388)*  weights = 30
            379) PERNP > 43000
              389)*  weights = 691
2) PINCP > 80070
  390) WKL == {3}; criterion = 1, statistic = 683.733
    391) RETP <= 14400; criterion = 1, statistic = 85.418
      392) OIP <= 18000; criterion = 1, statistic = 51.383
        393) PINCP <= 88400; criterion = 0.988, statistic = 14.112
          394)*  weights = 19
        393) PINCP > 88400
          395)*  weights = 29
      392) OIP > 18000
        396)*  weights = 16
    391) RETP > 14400
      397) RETP <= 44000; criterion = 1, statistic = 21.608
        398) OIP <= 6900; criterion = 1, statistic = 25.281
          399) PINCP <= 111000; criterion = 0.998, statistic = 17.833
            400) SSP <= 18800; criterion = 0.991, statistic = 14.721
              401)*  weights = 13
            400) SSP > 18800
              402)*  weights = 18
          399) PINCP > 111000
            403)*  weights = 7
        398) OIP > 6900
          404)*  weights = 23
      397) RETP > 44000
        405) PINCP <= 196900; criterion = 1, statistic = 21.996
```

```
      406)*  weights = 193
    405) PINCP > 196900
      407)*  weights = 11
390) WKL == {1, 2}
  408) AGEP <= 68; criterion = 1, statistic = 182.467
    409) WAGP <= 75000; criterion = 1, statistic = 88.451
      410) PINCP <= 204000; criterion = 1, statistic = 32.394
        411) SEMP <= 10000; criterion = 0.998, statistic = 45.125
          412) RETP <= 2800; criterion = 1, statistic = 43.964
            413) PINCP <= 91600; criterion = 1, statistic = 43.724
              414) PERNP <= 51000; criterion = 0.998, statistic = 20.671
                415)*  weights = 12
              414) PERNP > 51000
                416) OIP <= 1700; criterion = 0.986, statistic = 26.468
                  417) HINS3 == {2}; criterion = 0.995, statistic = 18.735
                    418) PERNP <= 60800; criterion = 1, statistic = 25.946
                      419)*  weights = 9
                    418) PERNP > 60800
                      420) PERNP <= 73000; criterion = 0.965, statistic = 14.728
                        421)*  weights = 11
                      420) PERNP > 73000
                        422)*  weights = 12
                  417) HINS3 == {1}
                    423)*  weights = 12
                416) OIP > 1700
                  424)*  weights = 23
            413) PINCP > 91600
              425) OIP <= 1600; criterion = 1, statistic = 26.477
                426) PERNP <= 70000; criterion = 1, statistic = 20.918
                  427)*  weights = 31
                426) PERNP > 70000
                  428)*  weights = 13
              425) OIP > 1600
                429)*  weights = 15
          412) RETP > 2800
            430)*  weights = 200
        411) SEMP > 10000
          431) PERNP <= 79000; criterion = 0.997, statistic = 21.869
            432)*  weights = 31
          431) PERNP > 79000
            433) PINCP <= 158000; criterion = 1, statistic = 31.519
              434) PINCP <= 114000; criterion = 0.963, statistic = 40.515
                435)*  weights = 148
              434) PINCP > 114000
                436) PERNP <= 111000; criterion = 1, statistic = 26.926
                  437)*  weights = 13
```

436) PERNP > 111000
438)* weights = 66
433) PINCP > 158000
439)* weights = 15
410) PINCP > 204000
440)* weights = 8
409) WAGP > 75000
441) PINCP <= 160850; criterion = 1, statistic = 188.061
442) MIGPUMA == {00390, 01700, 55102}; criterion = 1, statistic = 161.88
443)* weights = 13
442) MIGPUMA == {00001, 00100, 00104, 00190, 00200, 00400, 00500, 00590, 00600, 00700, 00800, 00900, 01000, 01100, 01200, 01300, 01400, 01490, 01500, 01690, 01800, 01900, 02000, 02102, 02200, 02300, 02400, 02500, 02600, 02700, 03300, 03600, 03700, 04007, 04100, 08100, 08690, 09500, 10700, 11600}
444) PINCP <= 123010; criterion = 1, statistic = 87.658
445) PERNP <= 80000; criterion = 1, statistic = 32.48
446) PERNP <= 78000; criterion = 0.998, statistic = 20.409
447)* weights = 18
446) PERNP > 78000
448)* weights = 73
445) PERNP > 80000
449) PINCP <= 100200; criterion = 1, statistic = 36.967
450) NWAV == {1, 2, 3}; criterion = 1, statistic = 68.842
451)* weights = 25
450) NWAV == {5}
452) PERNP <= 90000; criterion = 1, statistic = 25.392
453) PINCP <= 92700; criterion = 1, statistic = 134.974
454) PERNP <= 84700; criterion = 0.997, statistic = 26.366
455) PINCP <= 84860; criterion = 1, statistic = 83.288
456)* weights = 194
455) PINCP > 84860
457)* weights = 19
454) PERNP > 84700
458) MSP == {1, 3, 6}; criterion = 0.998, statistic = 29.018
459) MAR == {2, 5}; criterion = 0.997, statistic = 26.652
460) SCHL == {22, 23}; criterion = 1, statistic = 53.019
461)* weights = 14
460) SCHL == {16, 17, 18, 19, 20, 21, 24}
462) POWPUMA == {00200, 00500, 00600, 00800, 01200, 01300, 01400, 01500, 01690, 01900, 02100, 02200, 02300}; criterion = 0.976, statistic = 42.622
463)* weights = 55
462) POWPUMA == {00100, 00390, 01100, 02500}
464)* weights = 12
459) MAR == {1}
465)* weights = 463
458) MSP == {2, 4, 5}

```
                    466) MAR == {3}; criterion = 0.98, statistic = 22.875
                      467)*  weights = 68
                    466) MAR == {1, 4}
                      468)*  weights = 7
                  453) PINCP > 92700
                    469)*  weights = 34
                452) PERNP > 90000
                  470)*  weights = 599
              449) PINCP > 100200
                471) PERNP <= 1e+05; criterion = 1, statistic = 104.32
                  472)*  weights = 114
                471) PERNP > 1e+05
                  473) PERNP <= 102000; criterion = 0.999, statistic = 56.583
                    474) PINCP <= 106500; criterion = 1, statistic = 38.708
                      475) PINCP <= 102200; criterion = 1, statistic = 33.373
                        476) COW == {3, 4, 5, 7}; criterion = 0.998, statistic = 28.953
                          477)*  weights = 10
                        476) COW == {1, 2}
                          478)*  weights = 35
                      475) PINCP > 102200
                        479)*  weights = 7
                    474) PINCP > 106500
                      480)*  weights = 10
                  473) PERNP > 102000
                    481) POWPUMA == {00100, 00590, 00800, 01300, 01900, 02300, 02400,
02500}; criterion = 1, statistic = 84.559
                      482)*  weights = 172
                    481) POWPUMA == {00001, 00104, 00190, 00200, 00390, 00400, 00500,
00600, 00700, 00900, 01000, 01100, 01200, 01400, 01500, 01690, 01800, 02000, 02100, 02200,
02600, 03300, 03400, 05900, 08500, 55102}
                      483) JWTRNS == {01, 04, 09}; criterion = 0.997, statistic = 35.028
                        484)*  weights = 450
                      483) JWTRNS == {02, 05, 07, 10, 11, 12}
                        485) COW == {2, 3, 5, 6}; criterion = 1, statistic = 55.981
                          486)*  weights = 13
                        485) COW == {1, 4, 7, 8}
                          487) MSP == {1}; criterion = 1, statistic = 39.864
                            488) POWPUMA == {00390, 00700, 01000, 01100, 01400, 01500, 01690,
01800, 02000, 02100}; criterion = 0.99, statistic = 45.112
                              489)*  weights = 46
                            488) POWPUMA == {00600, 00900, 01200, 02200, 02600, 03400, 08500}
                              490)*  weights = 7
                          487) MSP == {2, 4, 6}
                            491)*  weights = 13
            444) PINCP > 123010
              492) PERNP <= 105000; criterion = 1, statistic = 171.335
```

```
            493) RETP <= 10300; criterion = 0.974, statistic = 16.362
              494)*  weights = 20
            493) RETP > 10300
              495)*  weights = 12
          492) PERNP > 105000
            496) PERNP <= 121000; criterion = 1, statistic = 52.89
              497)*  weights = 39
            496) PERNP > 121000
              498)*  weights = 680
        441) PINCP > 160850
          499) PERNP <= 160000; criterion = 1, statistic = 82.148
            500)*  weights = 70
          499) PERNP > 160000
            501) PINCP <= 219500; criterion = 1, statistic = 47.95
              502)*  weights = 339
            501) PINCP > 219500
              503) PERNP <= 205000; criterion = 1, statistic = 63.395
                504)*  weights = 18
              503) PERNP > 205000
                505) PERNP <= 223000; criterion = 0.986, statistic = 17.601
                  506) PINCP <= 222000; criterion = 0.968, statistic = 17.74
                    507)*  weights = 15
                  506) PINCP > 222000
                    508)*  weights = 11
                505) PERNP > 223000
                  509) MAR == {2, 3}; criterion = 0.999, statistic = 31.246
                    510)*  weights = 8
                  509) MAR == {1, 5}
                    511) OCCP == {4, 8, 18}; criterion = 0.999, statistic = 47.637
                      512)*  weights = 12
                    511) OCCP == {1, 2, 3, 5, 6, 9, 11, 12, 17, 20, 23, 24}
                      513)*  weights = 96
    408) AGEP > 68
      514)*  weights = 282
1) PINCP > 253000
  515) PERNP <= 190000; criterion = 1, statistic = 381.306
    516) RETP <= 80000; criterion = 1, statistic = 67.718
      517)*  weights = 163
    516) RETP > 80000
      518)*  weights = 15
  515) PERNP > 190000
    519) PINCP <= 682900; criterion = 1, statistic = 105.816
      520) PERNP <= 225000; criterion = 1, statistic = 34.134
        521)*  weights = 9
      520) PERNP > 225000
        522) PINCP <= 513500; criterion = 0.967, statistic = 45.324
```

```
      523) PERNP <= 256000; criterion = 1, statistic = 35.253
        524)*  weights = 23
      523) PERNP > 256000
        525) MAR == {3, 4}; criterion = 1, statistic = 33.229
          526)*  weights = 27
        525) MAR == {1, 2, 5}
          527)*  weights = 308
    522) PINCP > 513500
      528) PERNP <= 476000; criterion = 1, statistic = 22.136
        529)*  weights = 16
      528) PERNP > 476000
        530)*  weights = 13
519) PINCP > 682900
  531)*  weights = 19
```

Text of CTREE classification tree

Conditional inference tree with 76 terminal nodes

Response: FINTP
Inputs: CIT, COW, DDRS, DEAR, DEYE, DOUT, DPHY, DRAT, DRATX, DREM, ENG, FER, GCL, GCM, GCR, HIMRKS, HINS1, HINS2, HINS3, HINS4, HINS5, HINS6, HINS7, JWTRNS, LANX, MAR, MARHD, MARHM, MARHT, MARHW, MIG, MIL, MLPA, MLPB, MLPCD, MLPE, MLPFG, MLPH, MLPI, MLPJ, MLPK, NWAB, NWAV, NWLA, NWLK, NWRE, RELSHIPP, SCH, SCHG, SCHL, SEX, WKL, WRK, ANC, DIS, ESP, ESR, HICOV, HISP, MIGPUMA, MSP, NATIVITY, NOP, OC, OCCP, PAOC, POWPUMA, PRIVCOV, PUBCOV, QTRBIR, RAC1P, RACAIAN, RACASN, RACBLK, RACNH, RACPI, RACSOR, RACWHT, RC, SFN, SFR, VPS, WAOB, PWGTP, AGEP, CITWP, INTP, JWMNP, JWRIP, MARHYP, OIP, PAP, RETP, SEMP, SSIP, SSP, WAGP, WKHP, WKWN, YOEP, JWAP, JWDP, PERNP, PINCP, POVPIP, RACNUM
Number of observations: 46693

1) ANC == {4}; criterion = 1, statistic = 8941.419
  2) NWLK == {3}; criterion = 1, statistic = 704.992
    3) ESR == {3, 6}; criterion = 1, statistic = 550.638
      4) RELSHIPP == {37}; criterion = 1, statistic = 248.485
        5) SCHL == {01, 05, 14, 18, 19, 21, 22, 23}; criterion = 1, statistic = 63.23
          6) MIGPUMA == {00200, 00700, 01100, 01300, 01400, 01900, 02000}; criterion = 0.994, statistic = 37.811
            7)* weights = 120
          6) MIGPUMA == {00600, 00800, 01000, 02100, 02500}
            8)* weights = 7
        5) SCHL == {11, 12, 13, 16, 17, 20}
          9) SCH == {2}; criterion = 0.998, statistic = 42.044
            10)* weights = 16
          9) SCH == {1}
            11) OCCP == {9, 15, 16, 17, 22, 23, 24}; criterion = 0.978, statistic = 36.9
              12) WKL == {1, 2}; criterion = 0.997, statistic = 37.878
                13)* weights = 42
              12) WKL == {3}
                14)* weights = 257
            11) OCCP == {1, 14, 19, 20}
              15)* weights = 16
      4) RELSHIPP == {20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 36, 38}
        16) NWRE == {1, 2}; criterion = 1, statistic = 225.078
          17)* weights = 27
        16) NWRE == {3}
          18) NWAV == {1, 2, 3}; criterion = 1, statistic = 64.377
            19)* weights = 13
          18) NWAV == {5}
            20) NWAB == {3}; criterion = 1, statistic = 31.104
              21) AGEP <= 61; criterion = 0.985, statistic = 24.494

```
          22) ENG == {2}; criterion = 0.999, statistic = 44.789
            23)*  weights = 13
          22) ENG == {1, 3, 4}
            24)*  weights = 467
        21) AGEP > 61
          25)*  weights = 586
      20) NWAB == {1, 2}
        26)*  weights = 21
  3) ESR == {1, 2, 4}
    27) RELSHIPP == {20, 21, 22, 23, 24, 26, 27, 29, 35, 37}; criterion = 1, statistic = 142.624
      28) RACWHT == {0}; criterion = 1, statistic = 74.701
        29)*  weights = 184
      28) RACWHT == {1}
        30) AGEP <= 49; criterion = 1, statistic = 67.961
          31)*  weights = 1792
        30) AGEP > 49
          32) MLPCD == {1}; criterion = 0.979, statistic = 26.053
            33)*  weights = 36
          32) MLPCD == {0}
            34)*  weights = 955
    27) RELSHIPP == {25, 28, 30, 31, 32, 33, 34, 36, 38}
      35) DOUT == {1}; criterion = 1, statistic = 79.653
        36) PUBCOV == {1}; criterion = 0.991, statistic = 30.856
          37) JWAP <= 110; criterion = 0.988, statistic = 28.875
            38) JWTRNS == {10, 12}; criterion = 1, statistic = 36.157
              39)*  weights = 8
            38) JWTRNS == {01, 02, 11}
              40)*  weights = 64
          37) JWAP > 110
            41)*  weights = 9
        36) PUBCOV == {2}
          42)*  weights = 9
      35) DOUT == {2}
        43) RACWHT == {0}; criterion = 0.999, statistic = 53.304
          44)*  weights = 103
        43) RACWHT == {1}
          45) RC == {1}; criterion = 0.996, statistic = 47.108
            46) JWDP <= 97; criterion = 0.977, statistic = 20.396
              47)*  weights = 167
            46) JWDP > 97
              48)*  weights = 27
          45) RC == {0}
            49)*  weights = 724
2) NWLK == {1, 2}
  50) AGEP <= 37; criterion = 1, statistic = 54.076
    51) WRK == {1}; criterion = 1, statistic = 34.334
```

```
      52)*  weights = 69
    51) WRK == {2}
      53) WAGP <= 22500; criterion = 1, statistic = 39.246
        54) NWAB == {3}; criterion = 0.974, statistic = 39.246
          55)*  weights = 14
        54) NWAB == {2}
          56)*  weights = 298
      53) WAGP > 22500
        57)*  weights = 8
  50) AGEP > 37
    58) DREM == {1}; criterion = 0.994, statistic = 24.378
      59) SCHL == {01, 11, 13, 14, 15, 16, 17, 18, 19, 20, 21}; criterion = 1, statistic = 56.56
        60)*  weights = 148
      59) SCHL == {12, 22, 24}
        61)*  weights = 14
    58) DREM == {2}
      62)*  weights = 909
1) ANC == {1, 2, 3}
  63) RELSHIPP == {20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 38};
criterion = 1, statistic = 504.409
    64) AGEP <= 75; criterion = 1, statistic = 300.189
      65) SCHL == {05, 06, 11, 12, 13, 14, 19, 20, 21, 22, 23, 24}; criterion = 1, statistic =
194.973
        66) NWRE == {1, 2}; criterion = 1, statistic = 112.348
          67)*  weights = 1028
        66) NWRE == {3}
          68) ANC == {2}; criterion = 1, statistic = 117.169
            69) NOP == {1, 2, 4, 6, 7}; criterion = 1, statistic = 83.721
              70) AGEP <= 37; criterion = 1, statistic = 84.036
                71)*  weights = 4143
              70) AGEP > 37
                72) OIP <= 18300; criterion = 0.984, statistic = 68.134
                  73) RELSHIPP == {21, 23, 24, 26, 28, 29, 30, 31, 33}; criterion = 0.988, statistic =
68.801
                    74) NWAV == {1, 2}; criterion = 0.998, statistic = 42.957
                      75)*  weights = 38
                    74) NWAV == {3, 5}
                      76)*  weights = 2267
                  73) RELSHIPP == {20, 22, 25, 27, 32, 34, 36, 38}
                    77) RELSHIPP == {27, 32, 34, 36, 38}; criterion = 0.997, statistic = 54.397
                      78) POVPIP <= 457; criterion = 0.998, statistic = 17.406
                        79)*  weights = 58
                      78) POVPIP > 457
                        80)*  weights = 10
                    77) RELSHIPP == {20, 22, 25}
                      81)*  weights = 4653
```

```
            72) OIP > 18300
              82)*  weights = 81
          69) NOP == {3, 5, 8}
            83) OIP <= 520; criterion = 0.994, statistic = 15.567
              84)*  weights = 63
            83) OIP > 520
              85)*  weights = 7
        68) ANC == {1, 3}
          86) RACBLK == {1}; criterion = 1, statistic = 110.082
            87) WKL == {2}; criterion = 1, statistic = 44.492
              88)*  weights = 27
            87) WKL == {1, 3}
              89)*  weights = 377
          86) RACBLK == {0}
            90) HINS3 == {2}; criterion = 1, statistic = 67.24
              91) WRK == {2}; criterion = 0.99, statistic = 62.232
                92) NWLA == {3}; criterion = 1, statistic = 54.312
                  93)*  weights = 17
                92) NWLA == {2}
                  94) ENG == {2, 3}; criterion = 0.994, statistic = 31.981
                    95)*  weights = 53
                  94) ENG == {1, 4}
                    96)*  weights = 1307
              91) WRK == {1}
                97)*  weights = 7705
            90) HINS3 == {1}
              98) NWAB == {1, 2}; criterion = 1, statistic = 30.19
                99)*  weights = 1380
              98) NWAB == {3}
                100) ESR == {1}; criterion = 1, statistic = 97.721
                  101)*  weights = 419
                100) ESR == {3, 6}
                  102)*  weights = 60
      65) SCHL == {01, 02, 03, 04, 07, 08, 09, 10, 15, 16, 17, 18}
        103) NWRE == {1, 3}; criterion = 1, statistic = 59.401
          104) RACBLK == {1}; criterion = 1, statistic = 61.126
            105)*  weights = 364
          104) RACBLK == {0}
            106) ANC == {1, 3}; criterion = 1, statistic = 63.015
              107) VPS == {06, 13}; criterion = 0.999, statistic = 47.937
                108) RELSHIPP == {21, 22, 23, 26, 27, 31, 32, 33, 34, 38}; criterion = 0.992, statistic
= 47.724
                  109)*  weights = 2692
                108) RELSHIPP == {20, 24, 25, 28, 29, 30, 36}
                  110)*  weights = 3332
              107) VPS == {01, 02, 03, 04, 05, 12}
```

```
          111)*  weights = 188
        106) ANC == {2}
         112) SCHL == {01, 02, 03, 07, 16, 17}; criterion = 1, statistic = 50.785
          113) SCHL == {01, 02, 03, 07, 17}; criterion = 1, statistic = 48.812
           114) SCHL == {01, 02, 03, 07}; criterion = 1, statistic = 35.861
            115)*  weights = 25
           114) SCHL == {17}
            116)*  weights = 320
          113) SCHL == {16}
           117) JWRIP <= 4; criterion = 1, statistic = 41.786
            118)*  weights = 2780
           117) JWRIP > 4
            119)*  weights = 23
         112) SCHL == {04, 09, 10, 15, 18}
          120)*  weights = 1307
    103) NWRE == {2}
     121)*  weights = 886
  64) AGEP > 75
   122) NWLK == {3}; criterion = 1, statistic = 312.614
    123) ESR == {1, 3}; criterion = 1, statistic = 72.847
     124)*  weights = 192
    123) ESR == {2, 6}
     125) NWLA == {3}; criterion = 1, statistic = 30.963
      126)*  weights = 259
     125) NWLA == {2}
      127)*  weights = 67
   122) NWLK == {1, 2}
    128) NWAB == {1, 2}; criterion = 1, statistic = 27.772
     129)*  weights = 2668
    128) NWAB == {3}
     130) NWRE == {1, 2}; criterion = 0.96, statistic = 19.095
      131)*  weights = 217
     130) NWRE == {3}
      132)*  weights = 39
63) RELSHIPP == {37}
  133) DDRS == {1}; criterion = 1, statistic = 47.117
   134) RACWHT == {1}; criterion = 0.995, statistic = 39.191
    135) DEYE == {2}; criterion = 0.98, statistic = 36.306
     136) MAR == {3, 5}; criterion = 1, statistic = 32.548
      137)*  weights = 46
     136) MAR == {1, 2}
      138)*  weights = 90
    135) DEYE == {1}
     139)*  weights = 36
   134) RACWHT == {0}
    140)*  weights = 24
```

```
133) DDRS == {2}
  141) RAC1P == {2, 5, 6}; criterion = 1, statistic = 53.302
    142)*  weights = 110
  141) RAC1P == {1, 3, 8, 9}
    143) MAR == {2, 4}; criterion = 1, statistic = 42.808
      144)*  weights = 32
    143) MAR == {1, 3, 5}
      145) NATIVITY == {2}; criterion = 1, statistic = 38.213
        146)*  weights = 9
      145) NATIVITY == {1}
        147) RAC1P == {3}; criterion = 1, statistic = 27.309
          148) HICOV == {2}; criterion = 0.963, statistic = 20
            149)*  weights = 9
          148) HICOV == {1}
            150)*  weights = 18
        147) RAC1P == {1, 8, 9}
          151)*  weights = 144
```

Test output of CFOREST

Random Forest using Conditional Inference Trees

Number of trees: 500

Response: INTP
Inputs: COW, FER, HIMRKS, HINS2, HINS3, MAR, MSP, AGEP, MARHYP, RETP, SSP, WAGP, PINCP, POVPIP
Number of observations: 30000

Text output of random forest (using MICE)

```
Call:
 randomForest(formula = as.formula(f), data = df_imputed1, keep.forest = TRUE)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 4

          Mean of squared residuals: 78219536
                    % Var explained: 77.02

Call:
 randomForest(formula = as.formula(f), data = df_imputed2, keep.forest = TRUE)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 4

          Mean of squared residuals: 81717979
                    % Var explained: 76.51

Call:
 randomForest(formula = as.formula(f), data = df_imputed3, keep.forest = TRUE)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 4

          Mean of squared residuals: 80881974
                    % Var explained: 77.15
Call:
 randomForest(formula = as.formula(f), data = df_imputed4, keep.forest = TRUE)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 4

          Mean of squared residuals: 83594843
                    % Var explained: 76.25
Call:
 randomForest(formula = as.formula(f), data = df_imputed5, keep.forest = TRUE)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 4

          Mean of squared residuals: 75033765
                    % Var explained: 77.52
```

R code

```{r **load package**}
library(haven)
library(sas7bdat)
library(tidyverse)
library(rpart)
library(rpart.plot)
library(party)
library(Amelia)
library(MASS)
library(nnet)
library(mice)
library(randomForest)
library(popbio)
```

```{r **read dataset**}
df_original = read.sas7bdat("psam_p27.sas7bdat")%>% na_if("")
df_original[df_original == "NaN"] = NA
df_original <- df_original[!is.na(df_original$INTP),] # Remove all respondents less than 15 years old
df_original$INTP[df_original$FINTP == 1] <- NA # set missing INTP as NA
```

```{r **combine one variable**}
# combine over 400 levels into 25 levels
df = df_original %>%
  mutate(OCCP = as.numeric(OCCP)) %>%
  mutate(OCCP = replace(OCCP, OCCP %in% c(0010:0440), 1),
         OCCP = replace(OCCP, OCCP %in% c(0500:0750), 2),
         OCCP = replace(OCCP, OCCP %in% c(0800:0960), 3),
         OCCP = replace(OCCP, OCCP %in% c(1005:1240), 4),
         OCCP = replace(OCCP, OCCP %in% c(1305:1560), 5),
         OCCP = replace(OCCP, OCCP %in% c(1600:1980), 6),
         OCCP = replace(OCCP, OCCP %in% c(2001:2060), 7),
         OCCP = replace(OCCP, OCCP %in% c(2100:2180), 8),
         OCCP = replace(OCCP, OCCP %in% c(2205:2555), 9),
         OCCP = replace(OCCP, OCCP %in% c(2600:2920), 10),
         OCCP = replace(OCCP, OCCP %in% c(3000:3550), 11),
         OCCP = replace(OCCP, OCCP %in% c(3601:3655), 12),
         OCCP = replace(OCCP, OCCP %in% c(3700:3960), 13),
         OCCP = replace(OCCP, OCCP %in% c(4000:4160), 14),
         OCCP = replace(OCCP, OCCP %in% c(4200:4255), 15),
         OCCP = replace(OCCP, OCCP %in% c(4330:4655), 16),
         OCCP = replace(OCCP, OCCP %in% c(4700:4965), 17),
         OCCP = replace(OCCP, OCCP %in% c(5000:5940), 18),
```

```
        OCCP = replace(OCCP, OCCP %in% c(6005:6130), 19),
        OCCP = replace(OCCP, OCCP %in% c(6200:6765), 20),
        OCCP = replace(OCCP, OCCP %in% c(6800:6950), 21),
        OCCP = replace(OCCP, OCCP %in% c(7000:7640), 22),
        OCCP = replace(OCCP, OCCP %in% c(7700:8990), 23),
        OCCP = replace(OCCP, OCCP %in% c(9005:9760), 24),
        OCCP = replace(OCCP, OCCP %in% c(9800:9830), 25))
```

```{r **create GUIDE description file**}
nvar <- ncol(df_original)
varnames <- names(df_original)
roles <- rep("s",nvar)
roles[varnames %in% varnames[c(11, 13:34, 38:44, 46:61, 64, 66:68, 70, 75, 77, 79, 83, 85:86,
89:90, 95, 97, 99:101, 103, 108, 110:113, 116:119, 121:124, 127:128, 130:131, 167)]] <- "c"
roles[varnames %in% varnames[c(1:4, 6:8, 125, 126, 132:166, 168:288, 5, 80:82, 84, 87:88, 91,
94, 96, 98, 106, 109, 114, 115, 129, 102)]] <- "x"
roles[varnames %in% varnames[c(10, 12, 36:37, 45, 62, 63, 65, 69, 71:74, 76, 78, 92, 93, 104,
105, 107, 120)]] <- "n"
roles[varnames %in% "INTP"] <- "d"
roles[varnames %in% "PWGTP"] = "w"
output = cbind(1:nvar,varnames,roles)
output = rbind(c(1, "", "x"), output)
write("data.txt",file="desc.txt")
write("NA",file="desc.txt",append=TRUE)
write("2",file="desc.txt",append=TRUE)
write.table(output,file="desc.txt", row.names=FALSE, col.names=FALSE, quote=FALSE,
append=TRUE)
```

```{r **clean dataset and set quantitative and qualitative variables**}
df_mustexcluded = df[, c(1:4, 6:8, 125, 126, 132:166, 168:288)]
df_toomany = df[, c(5, 80:82, 84, 87:88, 91, 94, 96, 98, 106, 109, 114, 115, 129)]
df_numeric = df[,c(9, 10, 12, 35:37, 45, 62, 63, 65, 69, 71:74, 76, 78, 92, 93, 104, 105, 107,
120)]
for (i in 1:ncol(df_numeric))
  df_numeric[,i] = as.numeric(df_numeric[,i])
df_cat = df[, c(11, 13:34, 38:44, 46:61, 64, 66:68, 70, 75, 77, 79, 83, 85:86, 89:90, 95, 97,
99:103, 108, 110:113, 116:119, 121:124, 127:128, 130:131, 167)]
for (i in 1:ncol(df_cat))
  df_cat[,i] = as.factor(df_cat[,i])
df_clean = cbind(df_cat, df_numeric)
y <- df_clean$INTP
w <- df_clean$PWGTP
miss <- is.na(y)
```

```
# the two methods to calculate the mean of INTP
imp = function(yhat)
  return((sum(w[!miss]*y[!miss])+sum(w[miss]*yhat[miss]))/sum(w))
ipw = function(p)
  return(sum(w[!miss]*y[!miss]/p[!miss])/sum(w[!miss]/p[!miss]))

# make a histogram to see how the distribution of INTP looks like
df_geom = df_clean %>%
  mutate(INTP = as.factor(INTP)) %>%
  count(INTP)
intp = c(-1300, 0, "0-10000", "10000-20000", "20000-30000", "30000-40000", "40000-50000",
"50000-60000", "60000-70000", "70000-80000", "80000-90000", 254000)
n = c(sum(df_geom$n[1]), sum(df_geom$n[2]), sum(df_geom$n[3:191]),
sum(df_geom$n[192:278]),
      sum(df_geom$n[279-338]), sum(df_geom$n[339-377]), sum(df_geom$n[378-414]),
      sum(df_geom$n[415-424]), sum(df_geom$n[425-434]), sum(df_geom$n[435-444]),
      sum(df_geom$n[445-450]), sum(df_geom$n[451]))
df_geom = as.data.frame(cbind(intp, n))
df_geom$n = as.numeric(df_geom$n)
df_geom %>% ggplot(mapping = aes(x = reorder(intp, n), y = n)) +
  geom_col() +  geom_text(aes(label = n), vjust = -0.5) +
  theme_classic() + ggtitle("Histogram of INTP") + xlab("INTP") + ylab("number of
observations")
```

```{r output GUIDE dataset}
df_guide = df
numeric = c(9, 10, 12, 35:37, 45, 62, 63, 65, 69, 71:74, 76, 78, 92, 93, 104, 105, 107, 120)
for(i in 1:288){
  if(i %in% numeric)
    df_guide[,i] = as.numeric(df_guide[,i])
else
    df_guide[,i] = as.factor(df_guide[,i])
}
FINTP_guide = rep("D", 46693)
FINTP_guide[df$INTP == 25400] = "T"
FINTP_guide[is.na(df$INTP)] = "C"
df_guide$FINTP = FINTP_guide
#write.csv(df_guide,file = "data.txt")
#write.csv(df_guide[sample(1:46693, 6000, replace = FALSE),],file = "sample.csv")
```

```{r rpart}
#the mean of INTP without any imputation
y_noimputation = sum(w[!miss]*df_clean$INTP[!miss])/sum(w[!miss])
cat("The mean without imputation is", y_noimputation, "\n")
```

```
a = Sys.time()
rt_reg = rpart(INTP~.-as.numeric(FINTP), weight = PWGTP, data = df_clean, method =
"anova")
yhat_rpart = predict(rt_reg, newdata = df_clean)
imp_rpart <- imp(yhat_rpart)
b = Sys.time()
cat("The mean with imputation using rpart is", imp_rpart, "+", b-a, "\n")

a = Sys.time()
rt_class = rpart(FINTP~.-INTP-PWGTP, data = df_clean, method = "class")
p_rpart <- predict(rt_class, newdata = df_clean, type = "prob")[, 1]
ipw_rpart <- ipw(p_rpart)
b = Sys.time()
cat("The mean with ipw using rpart is",ipw_rpart, "+", b-a)
```

```{r ctree}
# drop missing value of INTP when fitting regression CTREE
df_na = df_clean %>%
  drop_na(INTP)

a = Sys.time()
ct_reg = ctree(INTP~.-FINTP, data = df_na)
yhat_ctree = predict(ct_reg, newdata = df_clean)
imp_ct <- imp(yhat_ctree)
b = Sys.time()
cat("The mean with imputation using ctree is", imp_ct, "+", b-a, "\n")

a = Sys.time()
ct_class = ctree(FINTP~.-INTP-PWGTP, data = df_clean)
#plot(ct_class)
p_ctree = predict(ct_class, newdata = df_clean, type = "prob")
temp = c()
for (i in 1:length(p_ctree)){
  temp = c(temp, p_ctree[[i]][1])
}
p_ctree = temp
p1_ctree =p_ctree
p1_ctree[p1_ctree==0]=min(p_ctree[p_ctree!=0])/2
ipw_ct = ipw(p1_ctree)
b = Sys.time()
cat("The mean with ipw using ctree is", ipw_ct,"+", b-a, "\n")
```

```{r amelia regression}
# use amelia to impute the dataset for regression
```

```
imp_scr = read.table("imp_reg.txt", header = TRUE)
# imp score larger than 2.4
imp_scr = imp_scr[1:14,]
df_amelia = droplevels(df_clean)
df_amelia = df_amelia[, names(df_amelia) %in% imp_scr$Variable]
#####MAR, MSP highly related
#####HINS2, HIMRKS
df_amelia_reg = df_amelia[,-c(4, 6)]
```

```{r amelia logistic regression}
fintp_numeric = as.numeric(df_clean$FINTP) - 1

imp_scr = read.table("imp_clas.txt", header = TRUE)
imp_scr = imp_scr[c(1:16),]
df_amelia = droplevels(df_clean)
df_amelia = df_amelia[, names(df_amelia) %in% imp_scr$Variable]
#MAR, MSP correlated
#RELSHIPP more than 10 levels
df_amelia = df_amelia[, -c(2, 6)]
out_class = amelia(df_amelia, noms = c(1:10), p2s = 0)

glm_amelia  = glm(fintp_numeric~., data = out_class$imputations$imp1)

a = Sys.time()
p_ipw_log1 = predict(glm(fintp_numeric~., data = out_class$imputations$imp1, family =
"binomial"), out_class$imputations$imp1, type = "response")
b = Sys.time()
p_ipw_log2 = predict(glm(fintp_numeric~., data = out_class$imputations$imp2, family =
"binomial"), out_class$imputations$imp2, type = "response")
p_ipw_log3 = predict(glm(fintp_numeric~., data = out_class$imputations$imp3, family =
"binomial"), out_class$imputations$imp3, type = "response")
p_ipw_log4 = predict(glm(fintp_numeric~., data = out_class$imputations$imp4, family =
"binomial"), out_class$imputations$imp4, type = "response")
p_ipw_log5 = predict(glm(fintp_numeric~., data = out_class$imputations$imp5, family =
"binomial"), out_class$imputations$imp5, type = "response")
ipw_log_amelia = (ipw(1-p_ipw_log1) + ipw(1-p_ipw_log2) + ipw(1-p_ipw_log3) + ipw(1-
p_ipw_log4)
                  + ipw(1-p_ipw_log5))/5
cat("The ipw using logistic regression with amelia is", ipw_log_amelia, "+", b-a)
```

```{r mice logistic regression}
out_mice = mice(df_amelia)
glm_mice = glm(abs(fintp_numeric-1)~., data = complete(out_mice,1), family = "binomial")
a = Sys.time()
```

```
p_ipw_mice1 = predict(glm(abs(fintp_numeric-1)~., data = complete(out_mice,1), family =
"binomial"),
                      complete(out_mice,1), type = "response")
b = Sys.time()
p_ipw_mice2 = predict(glm(abs(fintp_numeric-1)~., data = complete(out_mice,2), family =
"binomial"),
                      complete(out_mice,2), type = "response")
p_ipw_mice3 = predict(glm(abs(fintp_numeric-1)~., data = complete(out_mice,3), family =
"binomial"),
                      complete(out_mice,3), type = "response")
p_ipw_mice4 = predict(glm(abs(fintp_numeric-1)~., data = complete(out_mice,4), family =
"binomial"),
                      complete(out_mice,4), type = "response")
p_ipw_mice5 = predict(glm(abs(fintp_numeric-1)~., data = complete(out_mice,5), family =
"binomial"),
                      complete(out_mice,5), type = "response")
ipw_log_mice = (ipw(p_ipw_mice1) + ipw(p_ipw_mice2) + ipw(p_ipw_mice3) +
ipw(p_ipw_mice4)+
                  ipw(p_ipw_mice5))/5
cat("The ipw using logistic regression with mice is", ipw_log_mice, "+", b-a)
```

```{r amelia random forest}
out = amelia(df_amelia_reg, noms = c(1:5), p2s = 0)

df_imputed = cbind(df_clean$INTP, as.data.frame(out$imputations$imp1))[!miss,]
colnames(df_imputed) =c("INTP", names(df_imputed[,-1]))
df_imputed1 = df_imputed[sample(x = 1:39656, size = 30000, replace = FALSE),]
df_imputed = cbind(df_clean$INTP, as.data.frame(out$imputations$imp2))[!miss,]
colnames(df_imputed) =c("INTP", names(df_imputed[,-1]))
df_imputed2 = df_imputed[sample(x = 1:39656, size = 30000, replace = FALSE),]
df_imputed = cbind(df_clean$INTP, as.data.frame(out$imputations$imp3))[!miss,]
colnames(df_imputed) =c("INTP", names(df_imputed[,-1]))
df_imputed3 = df_imputed[sample(x = 1:39656, size = 30000, replace = FALSE),]
df_imputed = cbind(df_clean$INTP, as.data.frame(out$imputations$imp4))[!miss,]
colnames(df_imputed) =c("INTP", names(df_imputed[,-1]))
df_imputed4 = df_imputed[sample(x = 1:39656, size = 30000, replace = FALSE),]
df_imputed = cbind(df_clean$INTP, as.data.frame(out$imputations$imp5))[!miss,]
colnames(df_imputed) =c("INTP", names(df_imputed[,-1]))
df_imputed5 = df_imputed[sample(x = 1:39656, size = 30000, replace = FALSE),]

f <- paste('INTP ~', paste(colnames(df_imputed1)[colnames(df_imputed1)!="INTP"],
      collapse = ' + '))

a = Sys.time()
rf1 = randomForest(as.formula(f), data=df_imputed1, keep.forest = TRUE)
```

```
yhat_rf1 = predict(rf1, newdata = out$imputations$imp1)
b = Sys.time()
rf2 = randomForest(as.formula(f), data=df_imputed2, keep.forest = TRUE)
yhat_rf2 = predict(rf2, newdata = out$imputations$imp2)
rf3 = randomForest(as.formula(f), data=df_imputed3, keep.forest = TRUE)
yhat_rf3 = predict(rf3, newdata = out$imputations$imp3)
rf4 = randomForest(as.formula(f), data=df_imputed4, keep.forest = TRUE)
yhat_rf4 = predict(rf4, newdata = out$imputations$imp4)
rf5 = randomForest(as.formula(f), data=df_imputed5, keep.forest = TRUE)
yhat_rf5 = predict(rf5, newdata = out$imputations$imp5)
(imp(yhat_rf5) + imp(yhat_rf4) + imp(yhat_rf3) + imp(yhat_rf2) + imp(yhat_rf1))/5
b-a
```

```{r cforest}
imp_scr = read.table("imp_reg.txt", header = TRUE)
# imp score larger than 2.4
imp_scr = imp_scr[1:14,]
df_amelia = droplevels(df_clean)
df_amelia = df_amelia[, names(df_amelia) %in% imp_scr$Variable]

df_cforest = as.data.frame(cbind(df_na$INTP, df_amelia[!is.na(df_clean$INTP),]))
colnames(df_cforest) = c("INTP", names(df_cforest[,-1]))

a = Sys.time()
cf = cforest(INTP~., data = df_cforest)
yhat_cf = predict(cf, newdata = df_amelia)
b = Sys.time()
imp(yhat_cf)
```

```{r mice randomforest}
out_mice_reg = mice(df_amelia_reg)

df_imputed = cbind(df_clean$INTP, as.data.frame(complete(out_mice_reg, 1)))[!miss,]
colnames(df_imputed) =c("INTP", names(df_imputed[,-1]))
df_imputed1 = df_imputed[sample(x = 1:39656, size = 30000, replace = FALSE),]
df_imputed = cbind(df_clean$INTP, as.data.frame(complete(out_mice_reg, 2)))[!miss,]
colnames(df_imputed) =c("INTP", names(df_imputed[,-1]))
df_imputed2 = df_imputed[sample(x = 1:39656, size = 30000, replace = FALSE),]
df_imputed = cbind(df_clean$INTP, as.data.frame(complete(out_mice_reg, 3)))[!miss,]
colnames(df_imputed) =c("INTP", names(df_imputed[,-1]))
df_imputed3 = df_imputed[sample(x = 1:39656, size = 30000, replace = FALSE),]
df_imputed = cbind(df_clean$INTP, as.data.frame(complete(out_mice_reg, 4)))[!miss,]
colnames(df_imputed) =c("INTP", names(df_imputed[,-1]))
df_imputed4 = df_imputed[sample(x = 1:39656, size = 30000, replace = FALSE),]
```

```
df_imputed = cbind(df_clean$INTP, as.data.frame(complete(out_mice_reg, 5)))[!miss,]
colnames(df_imputed) =c("INTP", names(df_imputed[,-1]))
df_imputed5 = df_imputed[sample(x = 1:39656, size = 30000, replace = FALSE),]

a = Sys.time()
rf1 = randomForest(as.formula(f), data=df_imputed1, keep.forest = TRUE)
yhat_rf1_mice = predict(rf1, newdata = complete(out_mice_reg, 1))
b = Sys.time()
rf2 = randomForest(as.formula(f), data=df_imputed2, keep.forest = TRUE)
yhat_rf2_mice = predict(rf2, newdata = complete(out_mice_reg, 2))
rf3 = randomForest(as.formula(f), data=df_imputed3, keep.forest = TRUE)
yhat_rf3_mice = predict(rf3, newdata = complete(out_mice_reg, 3))
rf4 = randomForest(as.formula(f), data=df_imputed4, keep.forest = TRUE)
yhat_rf4_mice = predict(rf4, newdata = complete(out_mice_reg, 4))
rf5 = randomForest(as.formula(f), data=df_imputed5, keep.forest = TRUE)
yhat_rf5_mice = predict(rf5, newdata = complete(out_mice_reg, 5))
(imp(yhat_rf5) + imp(yhat_rf4) + imp(yhat_rf3) + imp(yhat_rf2) + imp(yhat_rf1))/5
b-a
```

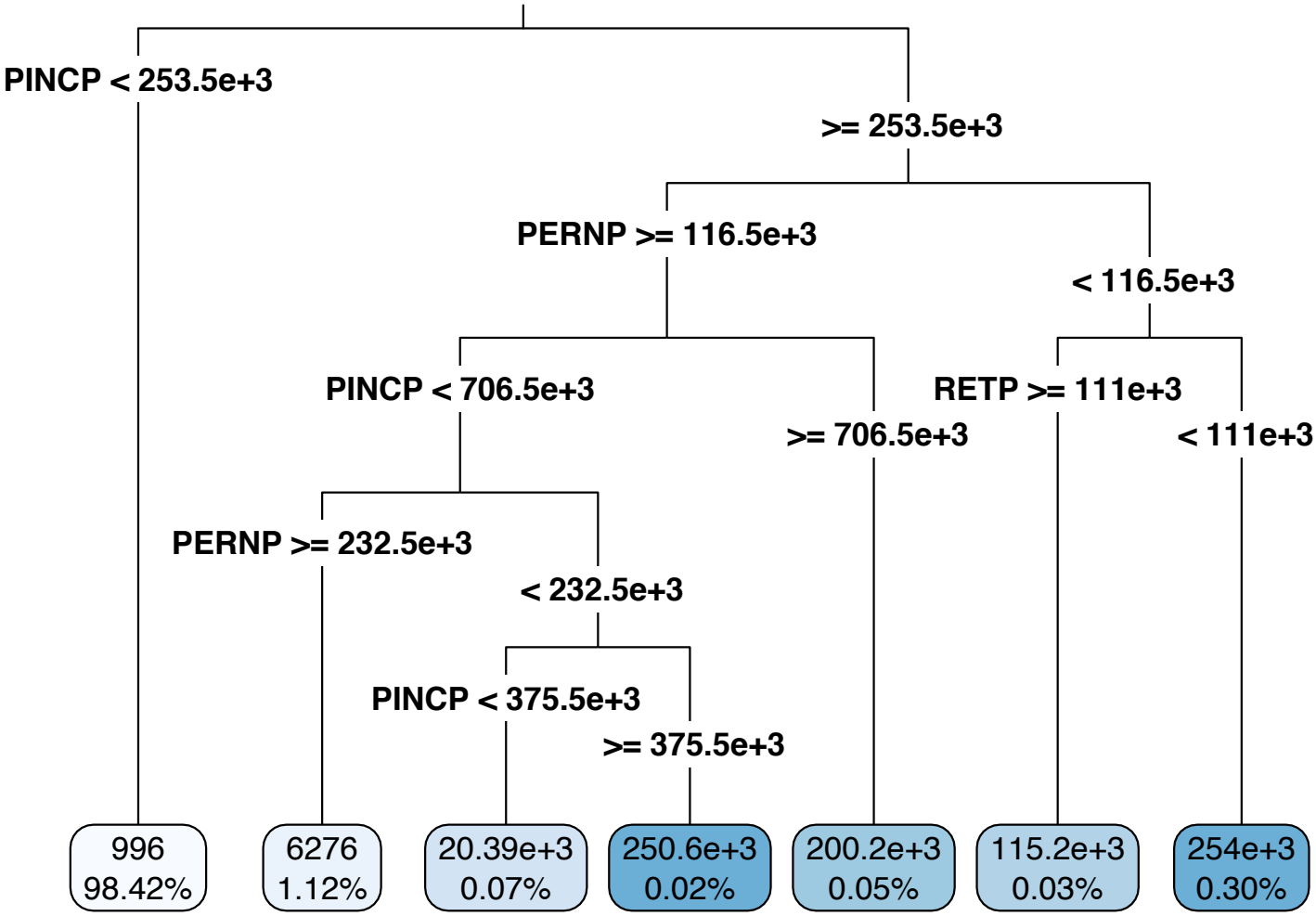
```{r guide estimated mean}
z <- df
group <- !is.na(z$INTP) ### group of nonmissing INTRDVX obs
w <- as.numeric(z$PWGTP) ### sampling weights

zreg <- read.table("fit_tree_reg.txt",header=TRUE)
imp(zreg$predicted)
yhat_guide_tree = zreg$predicted

zclass <- read.table("fit_tree_class.txt",header=TRUE)
ipw(zclass[,6])
p_guide_tree = zclass[,6]

zreg_forest = read.table("fit_forest_reg.txt",header=TRUE)
imp(zreg_forest$predicted)
yhat_guide_forest = zreg_forest$predicted
(zreg_forest[2] - zreg_forest[3])^2

zclass_forest <- read.table("fit_forest_class.txt",header=TRUE)
ipw(zclass_forest[,3])
p_guide_forest = zclass_forest[,3]

zclass_no <- read.table("fit_tree_class_no.txt",header=TRUE)
ipw(zclass_no[,6])
p_guide_tree_no = zclass_no[,6]
```

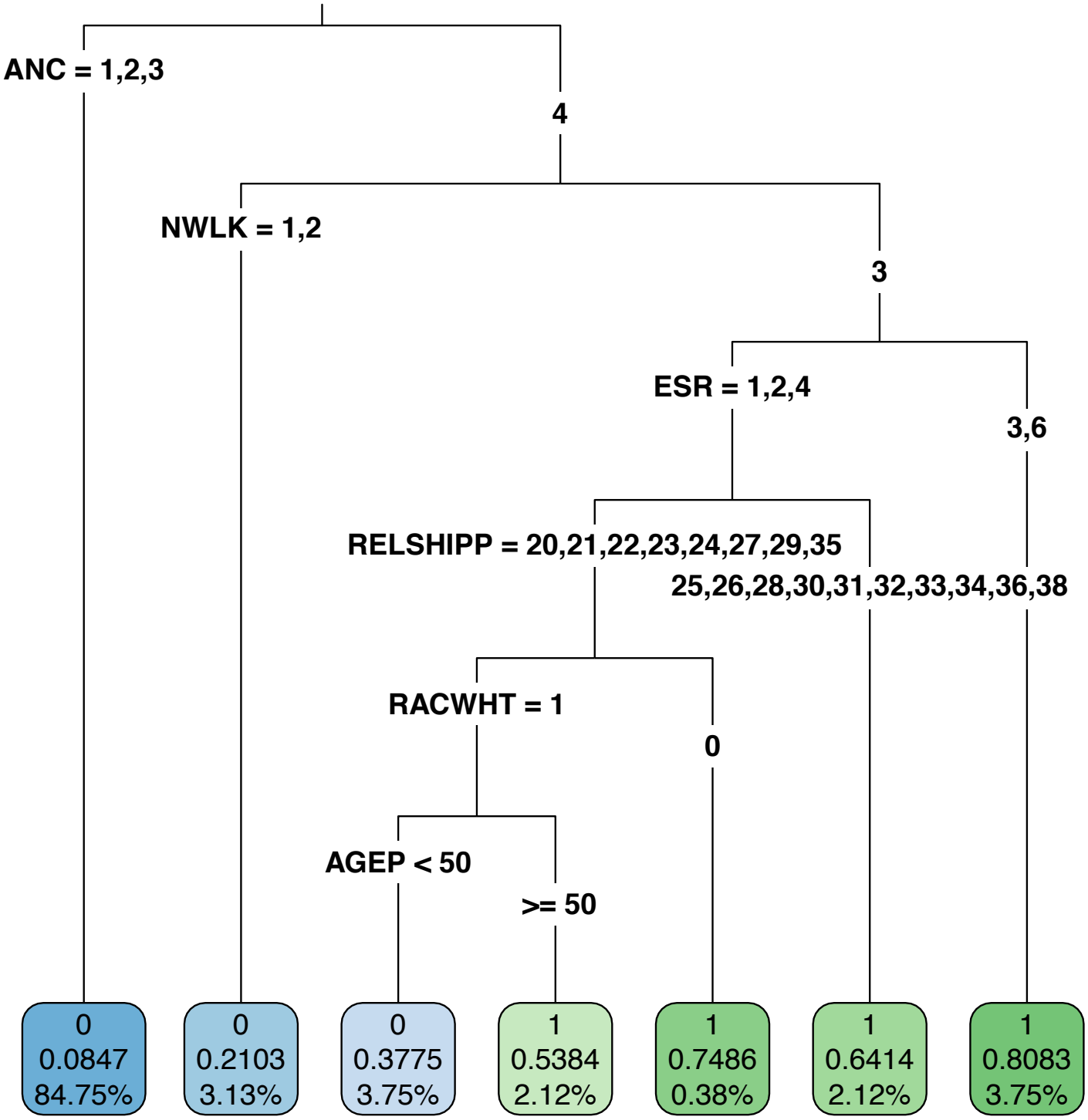
```
zclass_no2 <- read.table("fit_tree_class_no2.txt",header=TRUE)
ipw(zclass_no2[,6])
p_guide_tree_no2 = zclass_no2[,6]

zclass_forest <- read.table("fit_forest_class.txt",header=TRUE)
ipw(zclass_forest[,3])
p_guide_forest = zclass_forest[,3]

zreg2 <- read.table("fit_tree_reg2.txt",header=TRUE)
imp(zreg2$predicted)
yhat_guide_tree2 = zreg2$predicted

zreg3 <- read.table("fit_tree_stepwise.txt",header=TRUE)
imp(zreg3$predicted)
yhat_guide_tree_step = zreg3$predicted
```

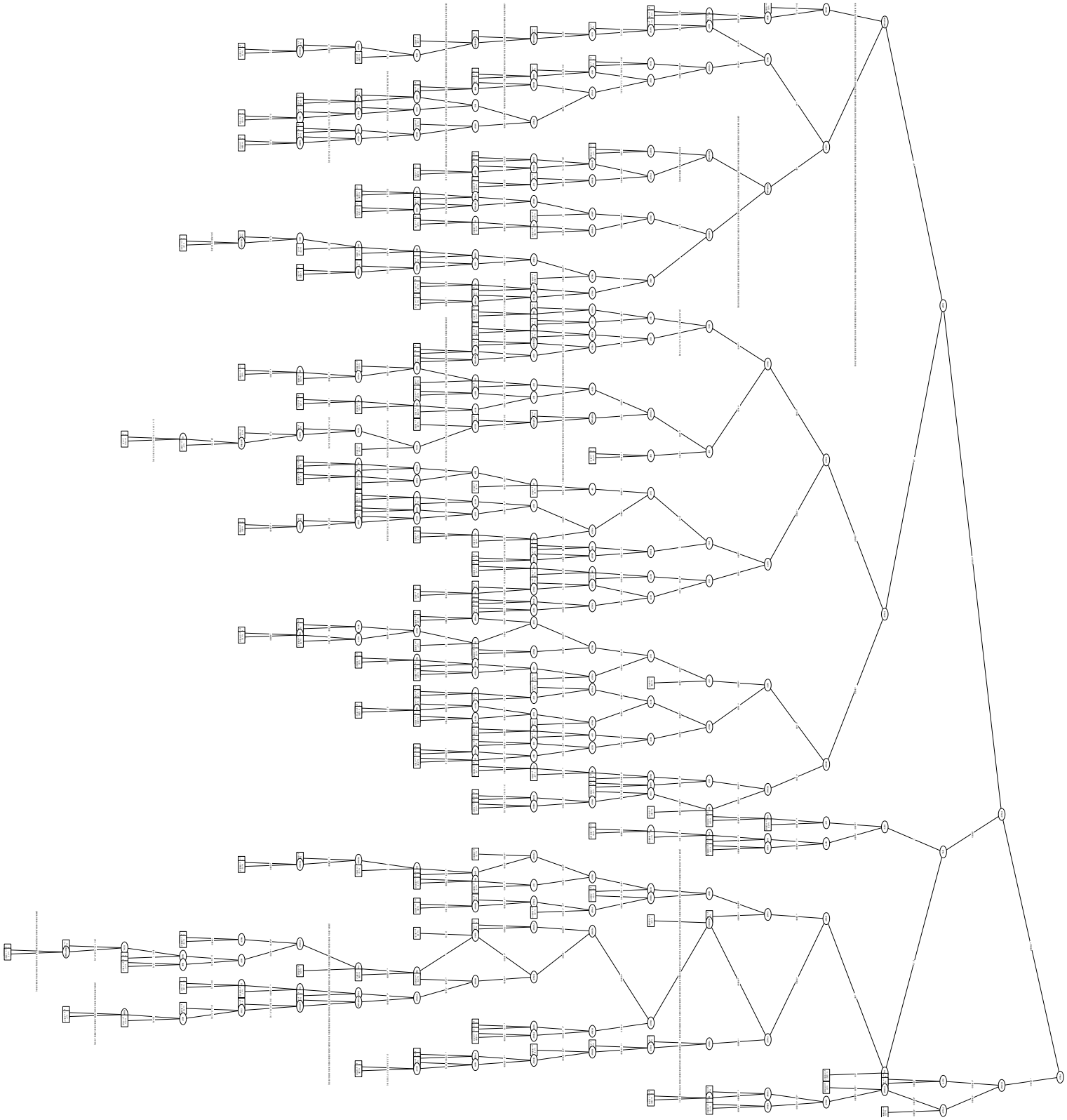
```{r important scoring regression}
pdf("imp_reg.pdf")
par(las=1,mar=c(5,12,4,2),cex=1)
leg.col <- c("orange","yellow")
leg.txt <- c("highly important","likely important")
x <- read.table("imp_reg.txt",header=TRUE)
score <- x$Score
vars <- x$Variable
type <- x$Type
barcol <- rep("orange",length(vars))
barcol[type == "L"] <- "yellow"
barcol[type == "U"] <- "cyan"
n <- sum(x$Type != "U")
barplot(rev(score[1:n]),names.arg=rev(vars[1:n]),
      col=rev(barcol[1:n]),horiz=TRUE,xlab="GUIDE importance scores", cex.names=0.5)
abline(v=1,col="red",lty=2)
legend("bottomright",legend=leg.txt,fill=leg.col)
dev.off()
```

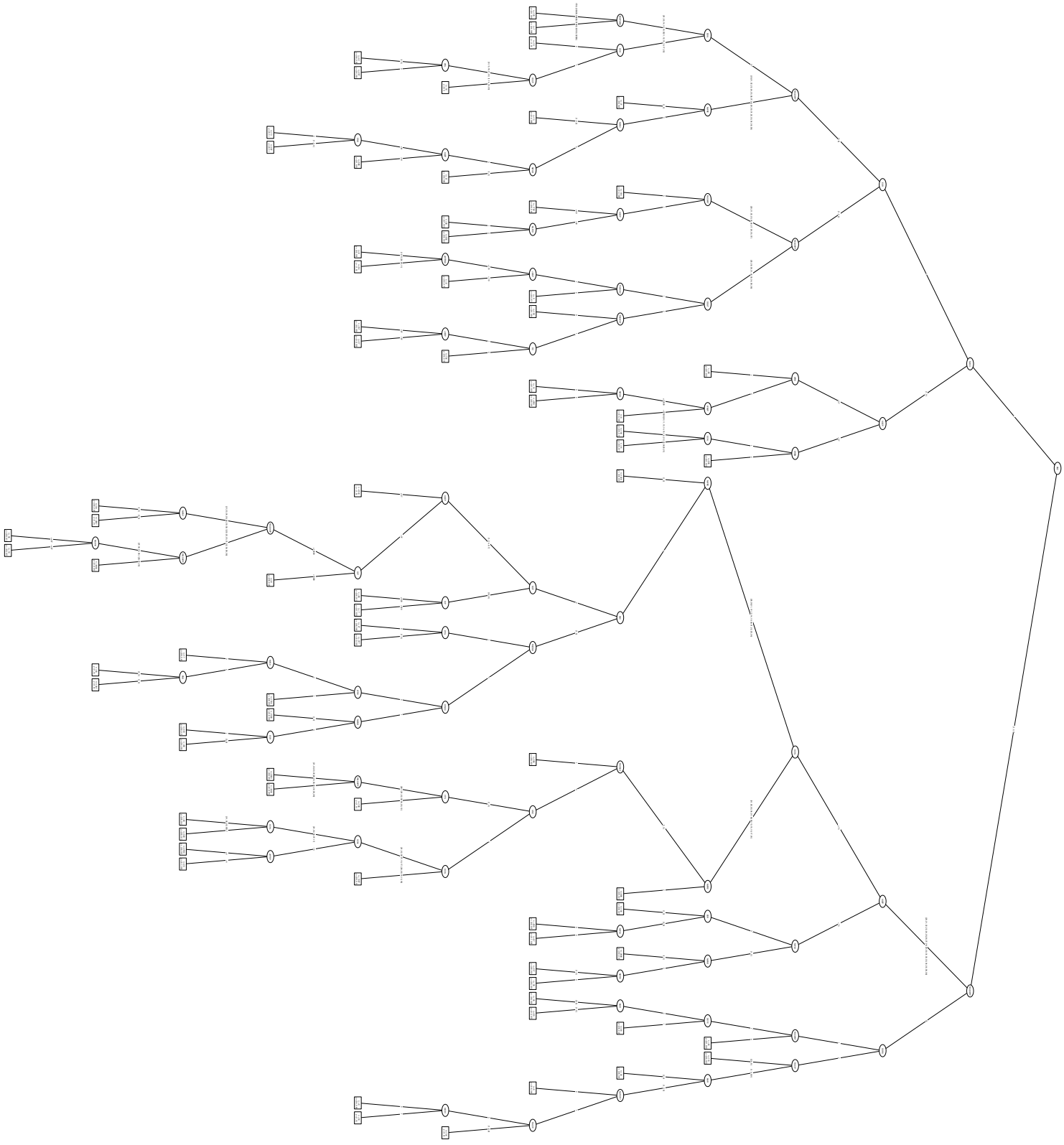
```{r important scoring classification}
pdf("imp_clas.pdf")
par(las=1,mar=c(5,12,4,2),cex=1)
leg.col <- c("orange","yellow")
leg.txt <- c("highly important","likely important")
x <- read.table("imp_clas.txt",header=TRUE)
score <- x$Score
vars <- x$Variable
```

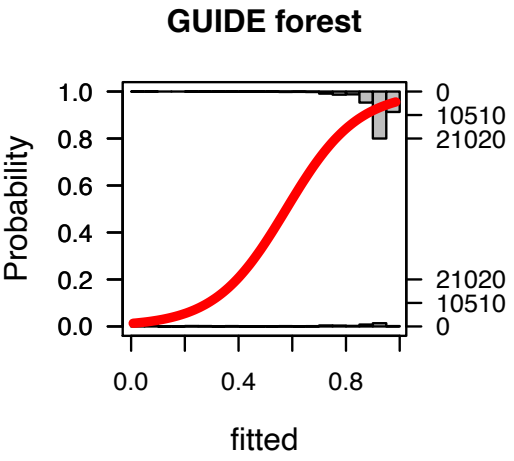
```
type <- x$Type
barcol <- rep("orange",length(vars))
barcol[type == "L"] <- "yellow"
barcol[type == "U"] <- "cyan"
n <- sum(x$Type != "U")
barplot(rev(score[1:n]),names.arg=rev(vars[1:n]),
      col=rev(barcol[1:n]),horiz=TRUE,xlab="GUIDE importance scores", cex.names=0.8)
abline(v=1,col="red",lty=2)
legend("bottomright",legend=leg.txt,fill=leg.col)
dev.off()
```

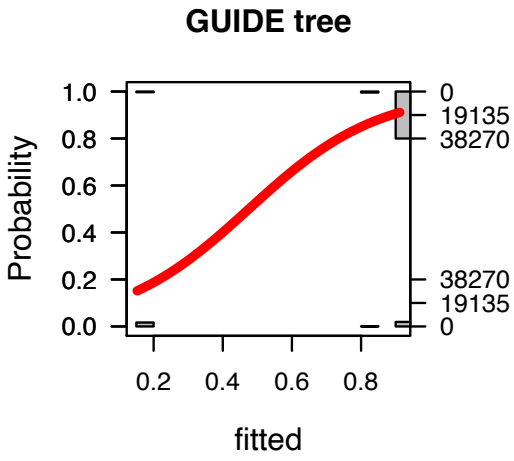

PINCP < 253.5e+3
>= 253.5e+3
PERNP >= 116.5e+3
< 116.5e+3
PINCP < 706.5e+3
>= 706.5e+3
RETP >= 111e+3
< 111e+3
PERNP >= 232.5e+3
< 232.5e+3
PINCP < 375.5e+3
>= 375.5e+3
996
98.42%
6276
1.12%
20.39e+3
0.07%
250.6e+3
0.02%
200.2e+3
0.05%
115.2e+3
0.03%
254e+3
0.30%

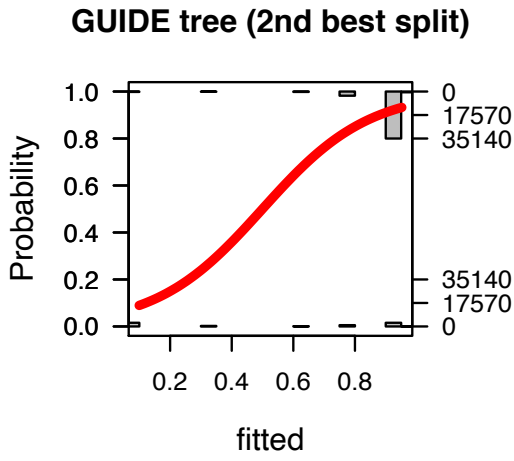

ANC = 1,2,3
4
NWLK = 1,2
3
ESR = 1,2,4
3,6
RELSHIPP = 20,21,22,23,24,27,29,35
25,26,28,30,31,32,33,34,36,38
RACWHT = 1
0
AGEP < 50
>= 50
0
0.0847
84.75%
0
0.2103
3.13%
0
0.3775
3.75%
1
0.5384
2.12%
1
0.7486
0.38%
1
0.6414
2.12%
1
0.8083
3.75%

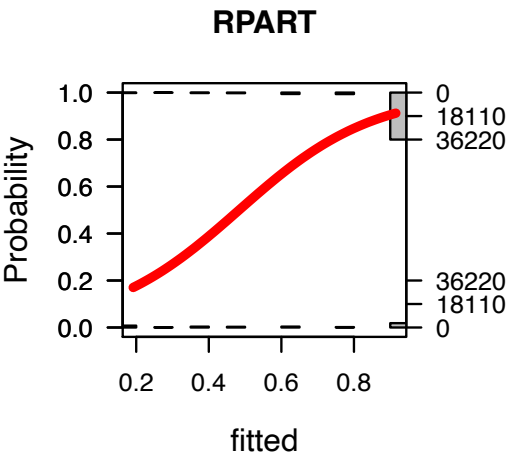


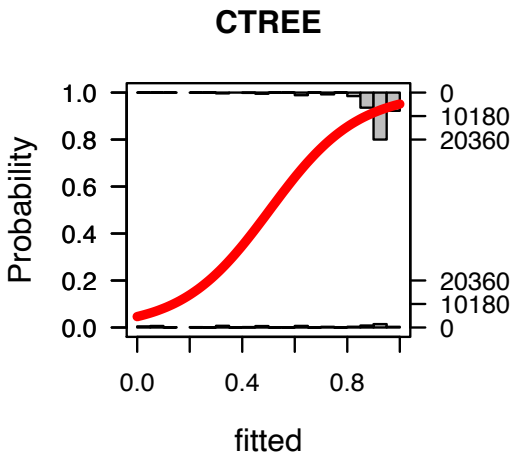


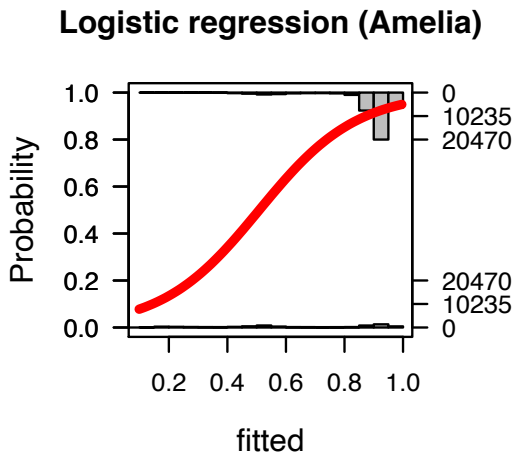

GUIDE forest
Probability
1.0
0.8
0.6
0.4
0.2
0.0
0
10510
21020
21020
10510
0
0.0
0.4
0.8
fitted

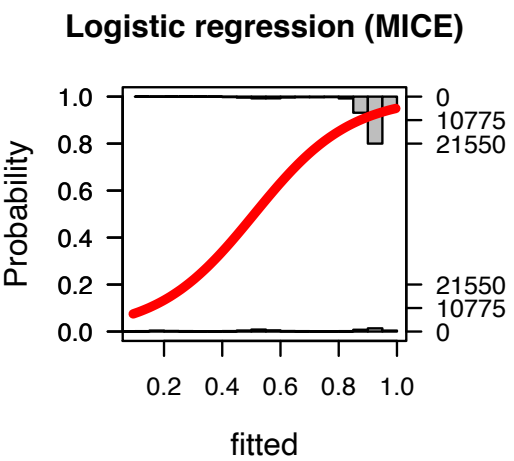

GUIDE tree
Probability
1.0
0.8
0.6
0.4
0.2
0.0
0
19135
38270
38270
19135
0
0.2
0.4
0.6
0.8
fitted

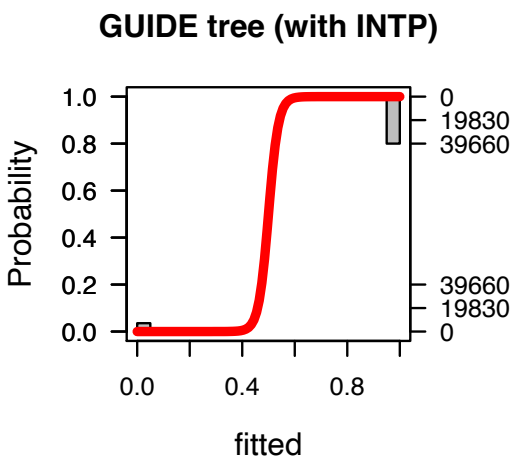

GUIDE tree (2nd best split)
Probability
1.0
0.8
0.6
0.4
0.2
0.0
0
17570
35140
35140
17570
0
0.2
0.4
0.6
0.8
fitted

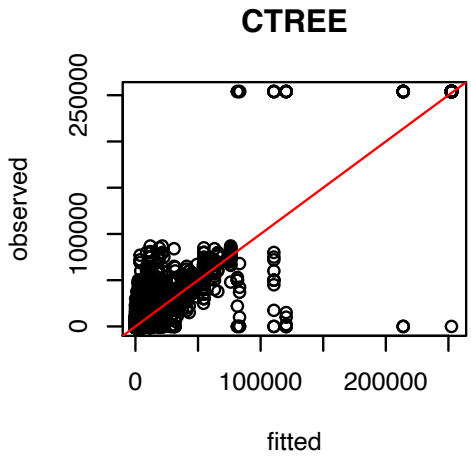

RPART
Probability
1.0
0.8
0.6
0.4
0.2
0.0
0
18110
36220
36220
18110
0
0.2
0.4
0.6
0.8
fitted

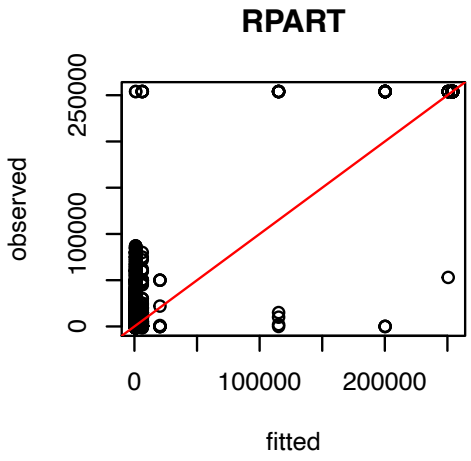

CTREE
Probability
1.0
0.8
0.6
0.4
0.2
0.0
0
10180
20360
20360
10180
0
0.0
0.4
0.8
fitted

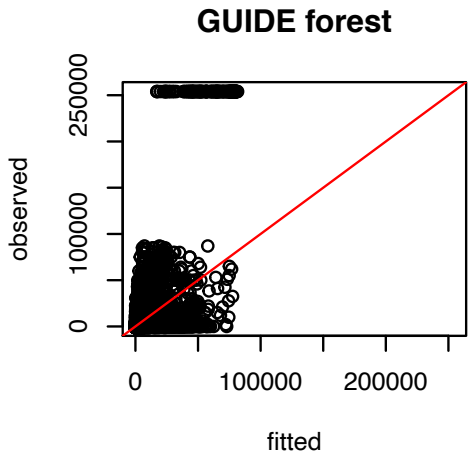

Logistic regression (Amelia)
Probability
1.0
0.8
0.6
0.4
0.2
0.0
0
10235
20470
20470
10235
0
0.2
0.4
0.6
0.8
1.0
fitted

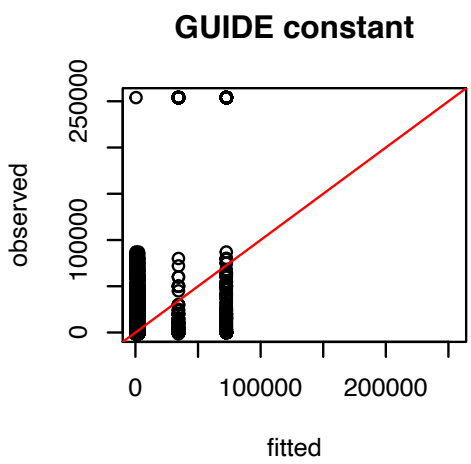

Logistic regression (MICE)
Probability
1.0
0.8
0.6
0.4
0.2
0.0
0
10775
21550
21550
10775
0
0.2
0.4
0.6
0.8
1.0
fitted

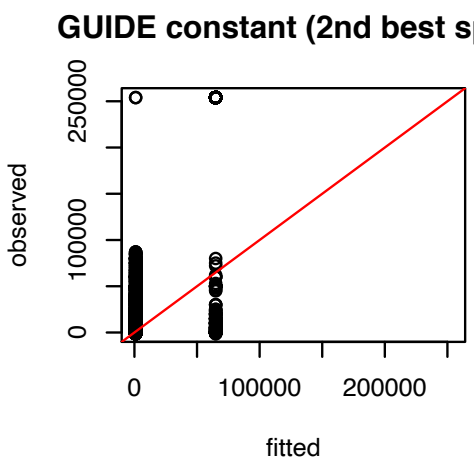

GUIDE tree (with INTP)
Probability
1.0
0.8
0.6
0.4
0.2
0.0
0
19830
39660
39660
19830
0
0.0
0.4
0.8
fitted

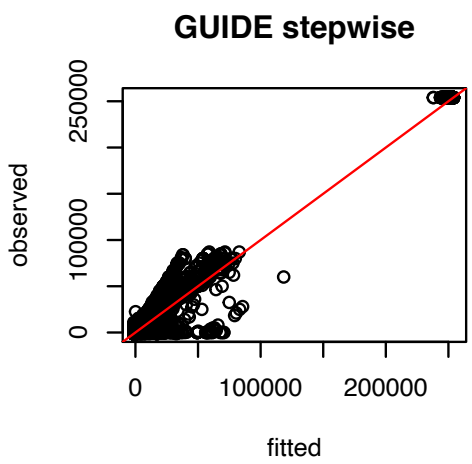

CTREE
observed
fitted
0
100000
200000
250000

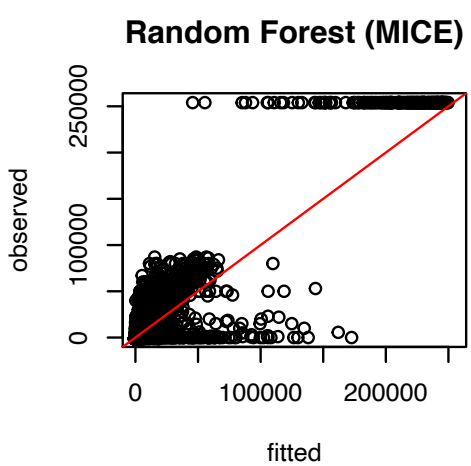

RPART
observed
fitted
0
100000
200000
250000

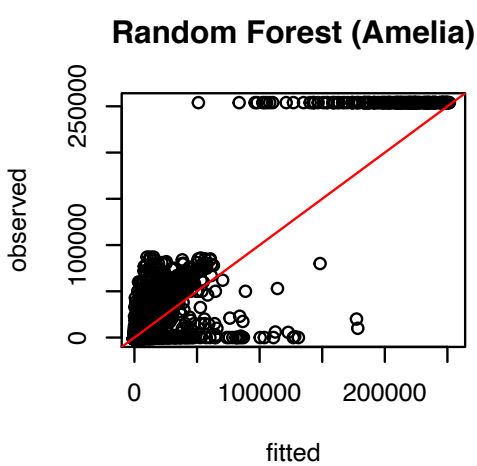

GUIDE forest
observed
fitted
0
100000
200000
250000

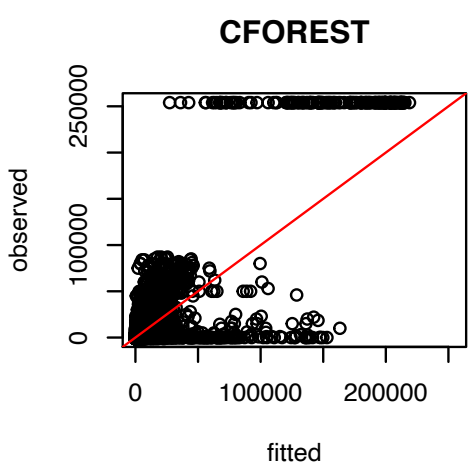

GUIDE constant
observed
fitted
0
100000
200000
250000

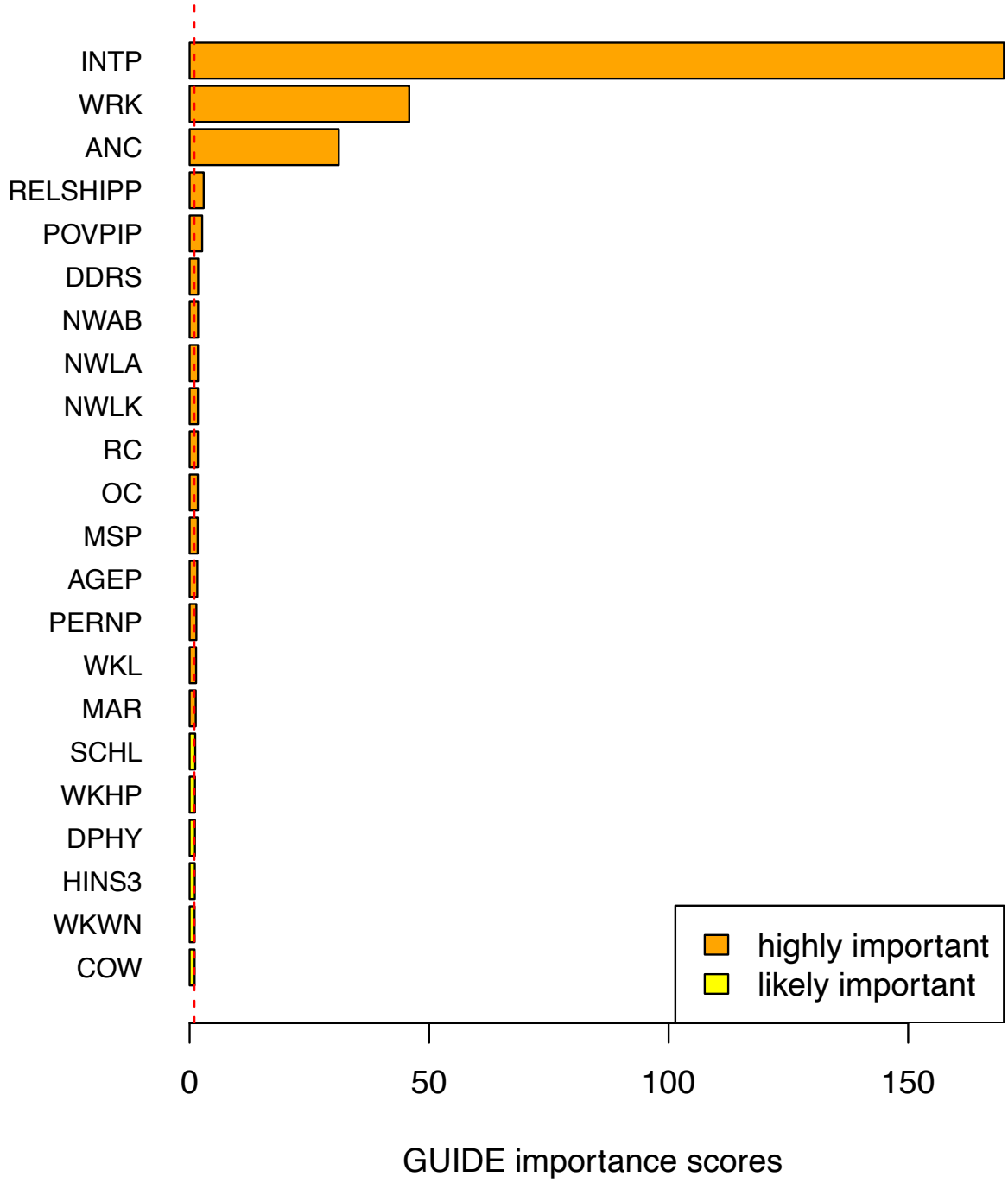

GUIDE constant (2nd best split)
observed
fitted
0
100000
200000
250000

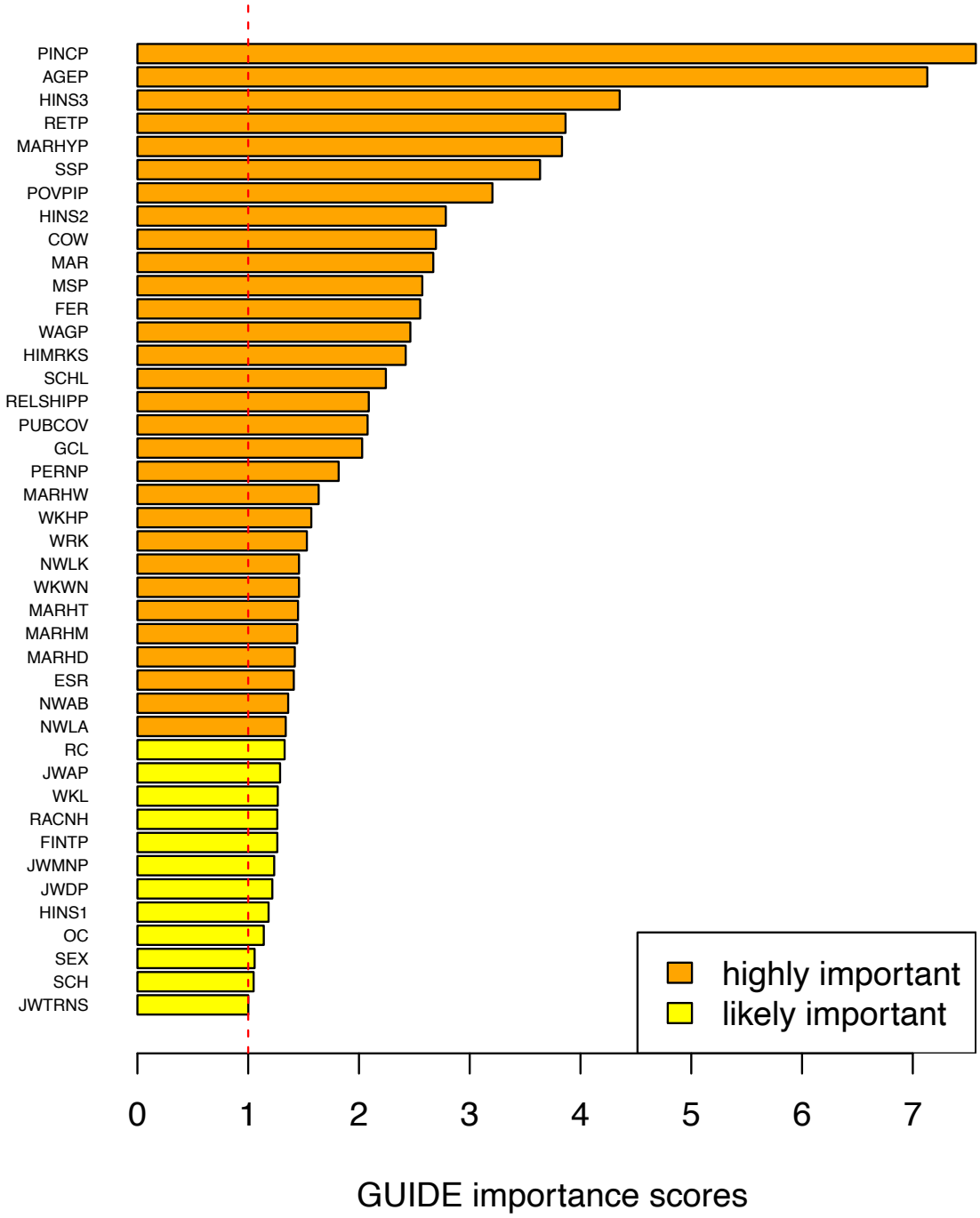

GUIDE stepwise
observed
fitted
0
100000
200000
250000


Random Forest (MICE)
observed
fitted
0
100000
200000
250000


Random Forest (Amelia)
observed
fitted
0
100000
200000
250000


CFOREST
observed
fitted
0
100000
200000
250000


INTP
WRK
ANC
RELSHIPP
POVPIP
DDRS
NWAB
NWLA
NWLK
RC
OC
MSP
AGEP
PERNP
WKL
MAR
SCHL
WKHP
DPHY
HINS3
WKWN
COW
highly important
likely important
0
50
100
150
GUIDE importance scores


PINCP
AGEP
HINS3
RETP
MARHYP
SSP
POVPIP
HINS2
COW
MAR
MSP
FER
WAGP
HIMRKS
SCHL
RELSHIPP
PUBCOV
GCL
PERNP
MARHW
WKHP
WRK
NWLK
WKWN
MARHT
MARHM
MARHD
ESR
NWAB
NWLA
RC
JWAP
WKL
RACNH
FINTP
JWMNP
JWDP
HINS1
OC
SEX
SCH
JWTRNS
highly important
likely important
0
1
2
3
4
5
6
7
GUIDE importance scores