Factors that influence labor participation in 1976

Xiao Tan

UIN: 676845530

NetID: xiaot6

ECON490 Machine Learning

May 15th, 2020

Table of Contents

Introduction

Labor force participation issue has always been a popular topic in Economics. According to Congressional Budget Office (2018), demographic factors can influence the labor force participation, including "sex, year of birth, education, marital status, and the presence of young children at home" (papa. 3). In addition, Jacobs (2015) suggests that policy, economic and fiscal factors also will influence the labor force participation. Also, Darian (1976) states that the presence of young kids would be the most important factor that influence women's labor force participation.

To analysis the factors that influence the labor force participation, this paper will explore the factors that influence labor participation in 1976, based on factors such as hours, age, kids, education, income, tax, college attendance in the family. We conducted this project to explore which factor would contribute most to the participation of labor in 1976, especially for the wives and family issues.

The variables I choose are more focused on wives and family factors. And here are some of my pre-analysis of the correlation between factors and labor participation, especially for the wives. Education might be related to the labor force for several reasons. Firstly, people with more education will be more competitive in the labor market, so they are more likely to get a job. Also, people with more education have a higher payment, which will further attract them to join the labor market. The effect of age is mixed since it will influence the labor market because the older age might choose to retire, while the older age might have more working experience. Tax will affect the labor market by affecting the incomes. The higher the tax is, the less likely the worker will join the working market. The experience will positively affect labor force participation because the more experience they have, it means they are more competitive in the

labor market. Also, the experience means they had once joined the labor market and are more likely to stay in the labor market. The young kids will negatively affect labor force participation because the young kids require a lot of care and attention, and the daycare and babysitting might be expensive for some family to afford, so the wives are more likely to stay at home to take care of young kids.

In this article, we will analysis the dataset from Online complements to Greene (2003). Table F4.1. And we will use several machine learning methods to analysis and predict the labor force participation at 1976.

Models and Analysis (Body)

1. Analysis of each variable.

1.1 Introduction of Variables:

In our dataset, the Y outcome is participation. Participation represents if the individual participates in the labor force in 1975. (This is essentially wage > 0 or hours > 0.) The participation is a binary variable, with "yes" and "no" for answers.

In our project, we choose the education, age, tax, experience, youngkids for main discussion.

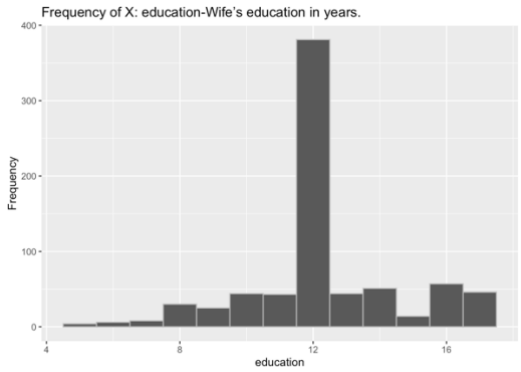
- education: Wife's education in years.
- age: Wife's age in years.
- wage: Wife's average hourly wage, in 1975 dollars.
- tax: Marginal tax rate facing the wife, and is taken from published federal tax tables (state and local
- experience: Actual years of wife's previous labor market experience.
- youngkids: Number of children less than 6 years old in household.

We will choose the education, age, tax, experience, and youngkids as our Xs to explain the Y. We want to analyze the labor force participant for wives, and we think the education, age, tax experience and youngkids would be the most influential factors intuitively.

1.2 Histogram of each variable

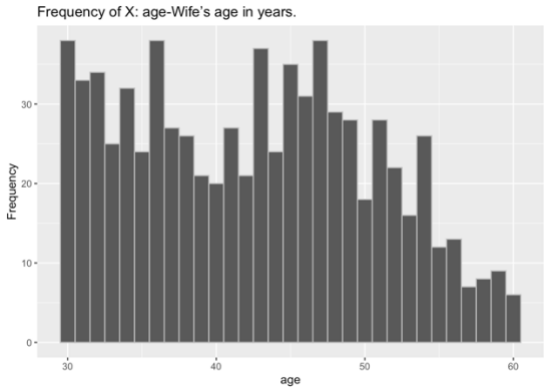
1.2.1 Education

Education represents the wife's education in years. The mode of education in the dataset is 12 years, which is probably high school graduation. And we have another increase in the frequency of education after that, 16 years, which is college graduation. This variable is similar to college attendance, so we only choose this one rather than bother of them. Also, we can see most of the wives have at least 8 years of education for most of the data points, and there is no data without any education.



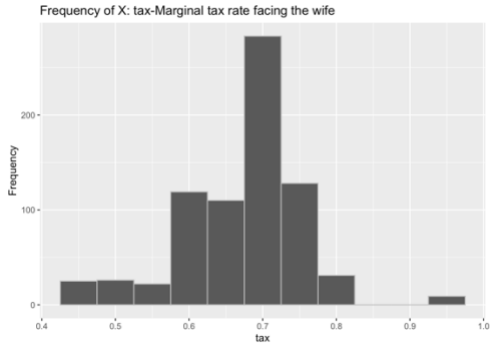
1.2.2 Age

Age here represents the wife's age in years. We can see that most of the data points have at least 30 years and at most 60 years old, and the evenly distributed in this range of 30-50 and a decrease after that. This range is good this age range is after graduation from school and before retirement age. And this is also the normal age for married women.



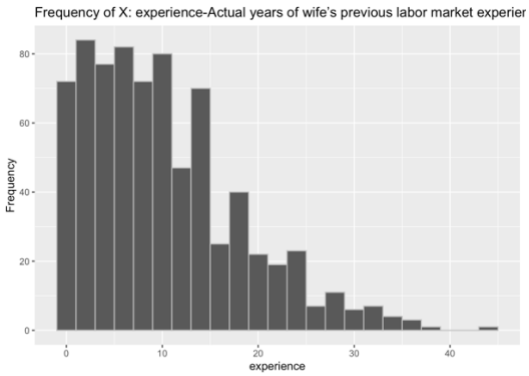
1.2.3 Tax

Tax here represents the marginal tax rate facing the wife and is taken from published federal tax tables (state and local). Most of the data points have a tax as more than 0.5 and the mode is around 0.7. This is probably the normal marginal tax rate of women in 1976.



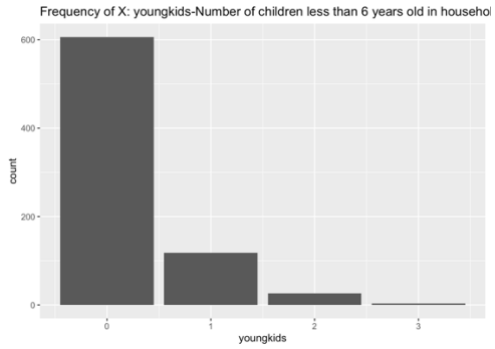
1.2.4 Experience

Experience here represents the actual years of the wife's previous labor market experience. We can see most of the wives have their experience range from 0 to 15 years. Also, there is a lot of them who have no working experience before. This is also caused by the normal age of working women.



1.2.5 Youngkids

Youngkids here represent the number of children less than 6 years old in the household. Most of the families in the dataset do not have young kids. And about 150 data points have young kids at home. Hence, the ratio of have young kids and do not have young kids is 4:1.

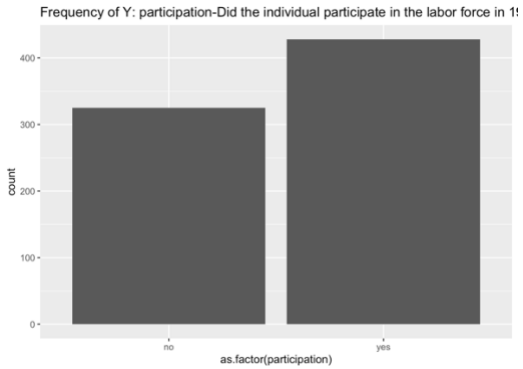


1.2.6 Wage

Wage here represent Wife's average hourly wage, in 1975 dollars. Most of the wife have the wages around 5 to 15 at 1975, and very few of them have more than 30.

1.2.7 Participation

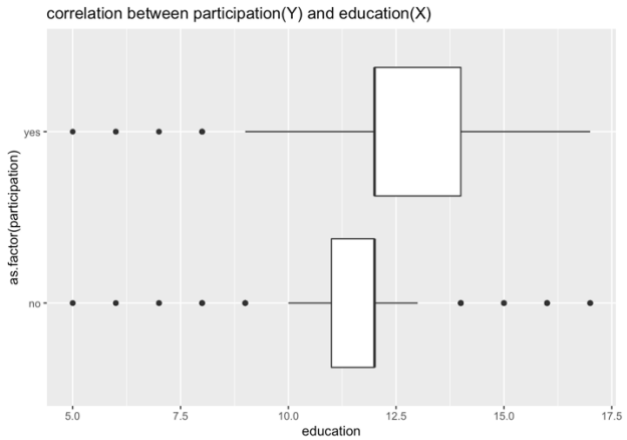
The participation results in our dataset are evenly distributed with a ratio of 7:9, which is close to 1:1. This is to say, the dataset is a fairly reasonable dataset with the output evenly distributed.



1.3 Correlation between Y and each selected X separately

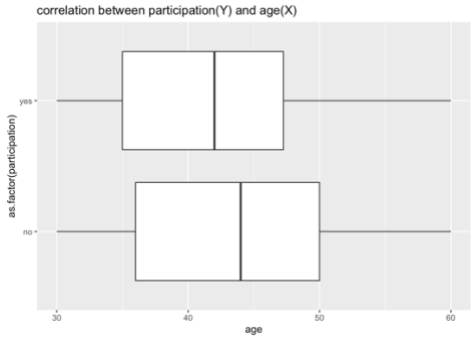
1.3.1 education

From the correlation between participation and education, we can see with more education, it is more likely to participate in the labor market in 1976. And differences between them are obvious because the 25-75 percentile of those two almost do not have a union. This means there is a strong correlation between education and participation.



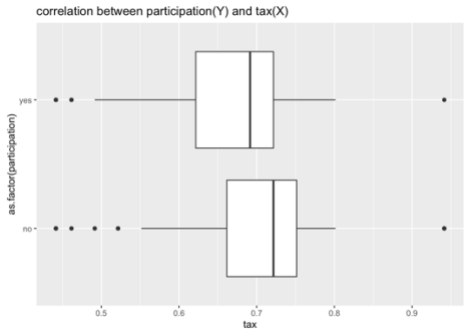
1.3.2 Age

From the correlation between participation and age, we can see the younger age wives are more likely to participate in the labor market in 1976, although the differences between those two are not very obvious, which means the age is not a good indicator of labor force participation for wives. This means there is a weak correlation between age and participation.



1.3.3 Tax

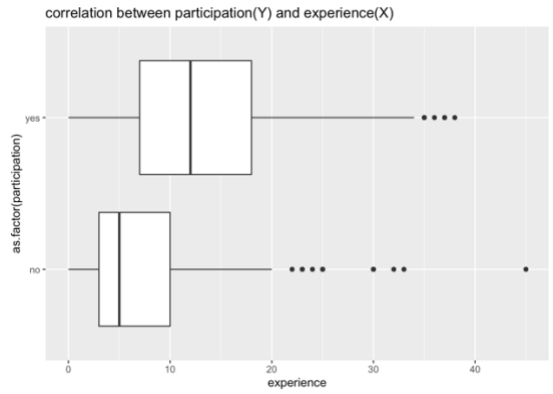
From the correlation of participation and tax, we can see with the lower the tax, it is more likely for the wives to participate in the labor market in 1976. This means there is a medium correlation between tax and participation.



1.3.4 Experience

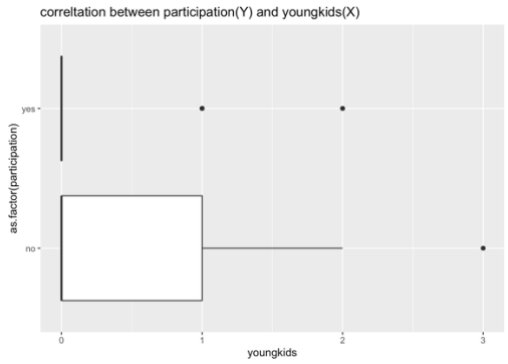
From the correlation of participation and experience, we can see with the more experience, it is more likely for the wives to participate in the labor market in 1976. And for

those who enter the labor market, seldom with zero working experience, and most of them have more than five years of working experience. This means there is a strong correlation between experience and participation.



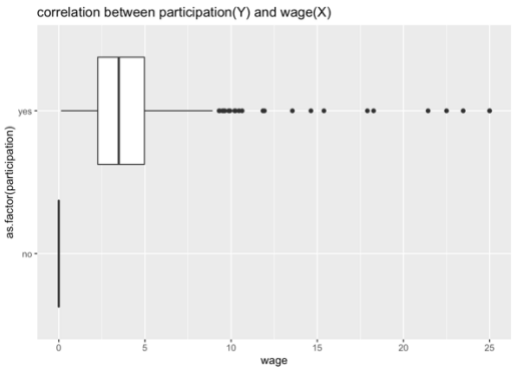
1.3.5 Youngkids

From the correlation of participation and youngkids, we can see with the existence of young kids, most of the women will not attend the labor market. This means there is a strong correlation between youngkids and participation.



1.3.6 Wage

From the correlation of participation and wage, we can see with wife will only attend the labor force market with wages.



2. Logistic classification

2.1 Introduction of Logistic classification

Compared with linear regression, logistic regression works better in the qualitative (categorical) examples. We estimate this probability using a logistic regression and then define a criterion to predict default for individuals with different balances. For instance, in our example, if the predicted probability is above 0.5, then we assign that as “yes” of participation; if the predicted probability is below 0.5, then we assign that as “no” of participation.

The objective function of logistic regression is:

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

In order to choose the best logistic classification model, we need to select the best model that have the highest prediction accuracy and also model interpretability, which means we need to drop the irreverent variables.

2.2 Backward stepwise selection

In this project, we use the backward stepwise selection. The advantage of backward stepwise selection is that we only need to go through 1 + p*(p + 1)/2 models. The disadvantages of backward stepwise selection are that we are not doing a thorough search compared with “best

subset selection and we need to have n > p observations since our benchmark model has all the regressors included.

By The backward stepwise selection comes with three steps:

- Step1: regression on all the regressors.
- Step2: drop the regressors one-at-a-time. Find the best p − 1 var regression call it Mp−1, then the best p − 2 var regression call it Mp−2, etc. You should find k best models in this step.
- Step3: choose the best model among M0, . . ., Mp by cross-validation

2.3 AIC and accuracy rate

And we compare the AIC and accuracy rate of all the variables.

The AIC is Akaike Information Criterion:

$$AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2)$$

A lower AIC value indicates a better fit.

The accuracy rate here is the number of correct predictions / all predictions.

2.4 Interpretation of our Logistic classification

Xs variables in the models	AIC value	accurancy rate
education+age+tax+experience+youngkids	AIC: 825.62;	0.7317397;
education+age+experience+youngkids	AIC: 828.94;	0.7383798;
age+experience+youngkids	AIC: 850.65;	0.7144754;
age+experience	AIC: 904.7;	0.6653386;
experience	AIC: 935.2;	0.6706507;
age	AIC: 1028.9;	0.5710491;
youngkids	AIC: 998.75;	0.622842;
education	AIC: 1006.7;	0.5909695;
tax	AIC: 1017.6;	0.5962815;

From the table, we can see that the AIC is increasing as we delete the variables from our full model, which means the full model has the highest accuracy. Also, from the accuracy rate, we can see that the full model and the model without tax have the highest accuracy rate. Combined with AIC, we will choose the full model as our optimal logistic regression model to analyze.

In this logistic regression model:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.69691    1.27953   2.889 0.003861 **
education    0.14744    0.04361   3.381 0.000723 ***
age         -0.10269    0.01338  -7.674 1.67e-14 ***
experience   0.12570    0.01328   9.467  < 2e-16 ***
tax         -2.62710    1.14926  -2.286 0.022259 *
youngkids   -1.41109    0.19810  -7.123 1.06e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the intercepts of the model, we can see that the absolute value of coefficients of the tax and youngkids are the largest, while the p value of tax is very large, so youngkids will have the largest effect on Y. The p-values of experience and age are the smallest, which means they are very significant.

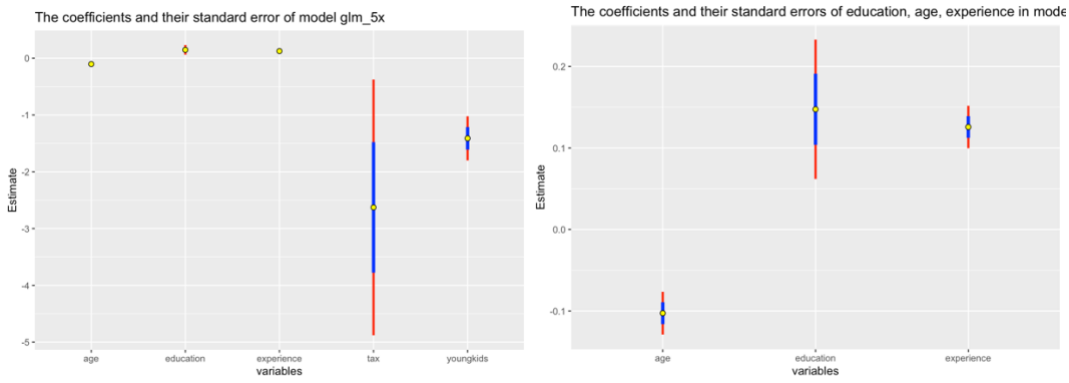
We can interpret the coefficients in this way:

- Holding other variables constant, if the education increases in one unit, which means the wife has one more year of education, the odds ratio of labor participation will increase at about 0.14 percent.
- Holding other variables constant, if the age increases in one unit (one year), which means the wife is one year older, the odds ratio of labor participation will decrease at about 0.10 percent.

- Holding other variables constant, if the tax increases in one unit, which is the marginal tax rate facing the wife, the odds ratio of labor participation will decrease at about 2.6 percent.
- Holding other variables constant, if the experience increases in one unit, which means the actual years of wife's previous labor market experience increase one year, the odds ratio of labor participation will increase at about 0.1257 percent.
- Holding other variables constant, if the youngkids increases in one unit, which means one more child less than 6 years old in the household, the odds ratio of labor participation will decrease at about 1.41 percent.

According to the p-value of each variable, education, age, experience, youngkids have p-values lower than 0.001, so they are significantly different from zero, with confidence level larger than 99%.

We can also check the standard error of each regressor:



We can see that the confidence interval of those variables does not contain 0. So, all the coefficients are significant.

Also, in terms of the accuracy rate, the true positive rate is 0.80, which means the model correctly predicts the positive class for 80.14019%. And the false positive rate is 0.36, which means the model incorrectly predicts the positive class for 36%.

This regression has a fairly good to predict labor participation in 1976 on the factors education, age, tax, experience, and youngkids.

3. Probit Classification

Probit model is very similar to the logit model. The difference is in the nonlinear function that we fit the data.

$$\text{Probit: } p(Y = 1|X) = p(X) = \Phi(\beta_0 + \beta_1 X)$$

The accuracy is 0.7343958 and the accuracy of the Logistic regression is 0.7317397, so it is slightly higher than Logistic regression. And the true positive rate is 0.7478261 and the true positive rate of Logistic regression is 0.7456522. Hence, it is also slightly higher than Logistic regression in terms of rate of true positive.

In general, the accuracy of probit regression and Logistic regression are very close. Probit regression's accuracy rate is slightly higher than the Logistic regression's accuracy rate.

4. Shrinkage methods

In shrinkage methods, we use all p of the regressors but use a technique that constraints, regularizes or shrinks the estimates. There are two well-known shrinkage methods called Ridge and LASSO (Least Absolute Selection and Shrinkage Operator).

4.1 Cross-Validation:

In both Ridge and LASSO, we need to use to cross-validation find the flexibly of the model (λ). Cross-validation is usually used to select the proper flexibility of a model or to evaluate its performance.

There are three methods of cross-validation:

- Set Approach: Divides the data to a train and a validation set (or hold-out set)
- Leave-One-Out Cross-Validation: Holds out only one observation in each period
- K-fold Cross-Validation: Holds out a subset of data in each of K periods

In our project, we use K-fold Cross-Validation (k = 10). The steps are:

- Step1: divide observations into k folds (groups) of equal size
- Step2: use the one of the folds as validation set and the remaining k − 1 folds as the training set
- Step3: calculate the MSE for the held-out fold
- Step4: calculate the k-fold CV estimate as:

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i$$

4.2 Ridge Regression

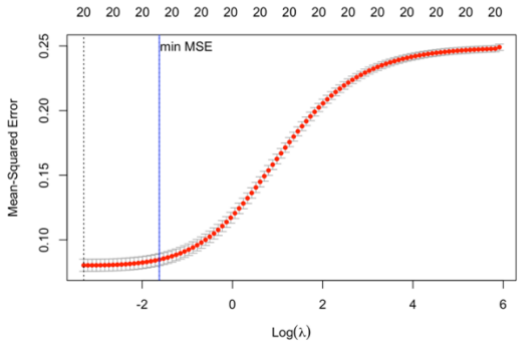
In Ridge regression, we minimize the following function:

$$\mathcal{L} = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = RSS + \underbrace{\lambda\sum_{j=1}^{p}\beta_j^2}_{\text{shrinkage penalty}}$$

Where λ is a tuning or hyper parameter and determines how strict the penalty or the constraint is. When λ = 0, ridge regression equals least squares regression. If λ = ∞, all coefficients are shrunk to zero.

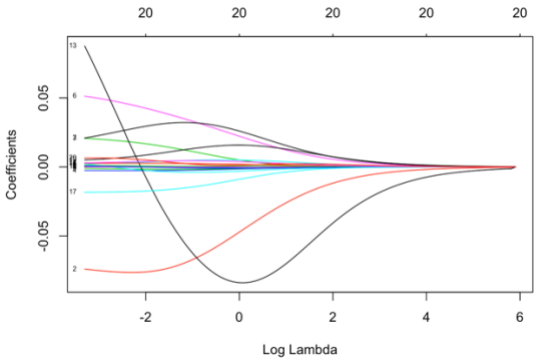
We use cross-validation to tune the model. The min value, λ = 0.03708, which is the one which minimizes out-of-sample loss in CV. The 1se value, λ= 0.19789, which is the largest λ within 1 standard error of the minimum λ.

We can plot the MSE as a function of λ:



We plot the MSE as a function of lambda. The number -1.7 on the top is the number of nonzero coefficients estimates. Confidence intervals represent error estimates for the loss metrics, which are the red dots. They are computed using cross validation. The left vertical line shows the location of the minimum lambda and the right vertical line shows the location of the lambda which is within 1 standard error of the minimum lambda.

We can also plot of how the coefficients vary with λ in a graph:



In this graph, we can see all coefficients converges to zero as λ larger than 3. When λ smaller than 1, the coefficients hasve some large change as shown above.

Finally, we use the λ = 0.19789, to run our ridge model, the accuracy rate of ridge model is 0.9309429.

Although ridge has computational advantages over other stepwise methods, it has one disadvantage. It shrinks the estimates towards zero (as one increases λ) but does not set any of

them exactly equal to zero. This is usually not a problem for prediction; however, it makes it hard to interpret the coefficients of interest.

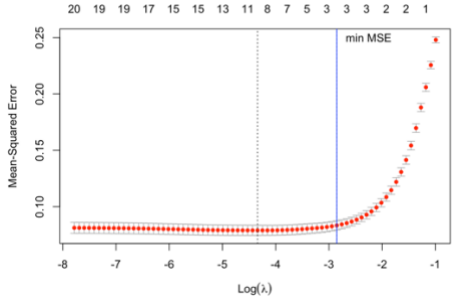
4.3 LASSO

Lasso is an improvement over ridge in variable selection dimension.

In Lasso regression, we minimize the following function:

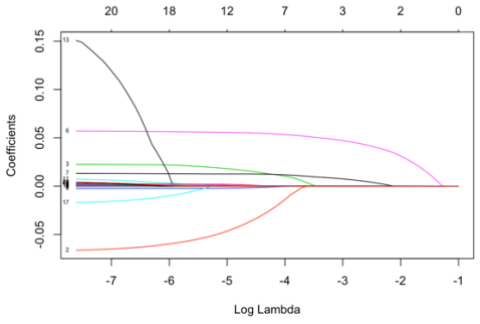
$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j| = RSS + \underbrace{\lambda\sum_{j=1}^{p}|\beta_j|}_{\text{shrinkage penalty}}$$

By plotting the MSE as a function of λ:



We can see that the min MSE is around log(λ) equals to -2.9.

By plotting how the coefficients vary with λ in a graph:



Each colored line represents the value taken by a different coefficient in the model. As λ grows, the regularization term has greater effect and there are fewer variables in the model

because more and more coefficients will be 0. By using the best $\lambda = 0.05768699$, the accuracy rate of lasso model is 0.9163347.

4.4 Comparing Ridge and LASSO

Lasso has an accuracy rate of 0.9163347 and Ridge has an accuracy rate of 0.9309429. We can see that ridge has a slightly higher accuracy rate. The accuracy of Lasso and Ridge are very close. None of them is universally dominant to the other.

Usually, Lasso performs better when only a smaller subset of variables is relevant for the model at hand. Ridge performs better when many of the predictors have prediction power. In general, lasso performs variable selection and results in smaller models which subsequently are easier to interpret.

5. KNN classification

KNN is a non-parametric method since we do not assume the functional form for the relation between Y and X.We specify K which is the number of nearest neighbors that we want to consider for each observation. The step of KNN is as follows:

- Step 1: Choose K
- Step 2: Find K neighbors for each observation (N0)
- Step3: Count neighbors in each group

$$Pr(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in N_0} I(y_i = j)$$

We also need to find the best K for the KNN classification. If K is too small the model is not smooth and it perfectly fits the data, but it does not perform well with new observation. This

is called Over Fitting. If K is too large the model is very smooth, but it does not perform well in the train and the test data. This is called Under Fitting.

To choose the best K, we can divide the data to the train and the test subsets, estimate the model using the train subset, test the performance of the model on the test subset and choose the K that performs best in the test subset.

After looping the K, we find the highest accuracy rate is 0.7869917 with k = 25.So, after predicting the results with on our entire data, we find the highest accuracy rate is 0.7869917 with k = 25.

Compared to the result with the Logistic regression, whose accuracy rate is 0.7317397, the KNN classification has a higher accuracy rate, but lower rate than LASSO and Ridge.

6. Decision tree

6.1 Full Decision trees.

The steps of stratifying X s and creating a regression tree:

- Step 1: Divide the predictor space into J mutually exclusive regions: R1, . . ., Rj.
- Step 2: Predict the outcome for all the observations that fall into Rj by calculating the average of the outcome for the training observations located in that region.

And we divide the predictor space into distinct and non-overlapping regions by recursive binary splitting.

- Start from the root (top). Find the best split minimizing the RSS. In other words, select Xj and a cut-off s that leads to the greatest reduction in RSS.

At this point, we have two boxes $R_1 \& R_2$:

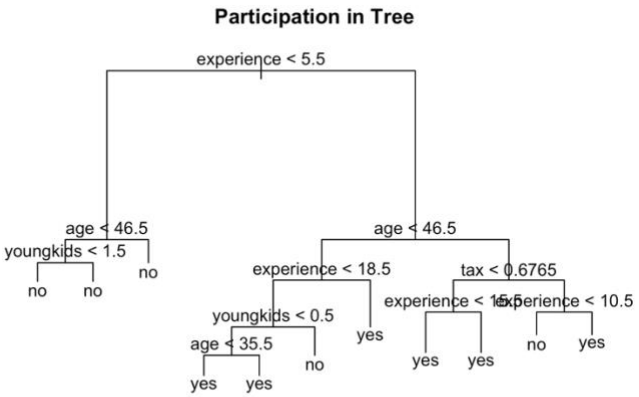
$$R_1 = X|X_j < s \text{ and } R_2 = X|X_j \geq s$$

s is selected to minimize $\sum_{i:x_i \in R_1}(y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2}(y_i - \hat{y}_{R_2})^2$

Now we have two new branches.

- Repeat step one at each internal node and split again to get 4 more branches.
- Keep going until your stopping criterion is reached!

By applying the above steps, we get the decision tree as follows:



This tree we have a height of 5 and accuracy rate of 0.6888889.

6.2 Prune tree

We consider whether pruning the tree might lead to improved results. To do cost complexity pruning we find a T that minimize the following RSS for a non-negative value of α:

$$\sum_{m=1}^{|T|} \sum_{i \in R_j} (y_i - \hat{y}_{R_m})^2 + \alpha |T|$$

R_m: the rectangle corresponding to mth terminal node.

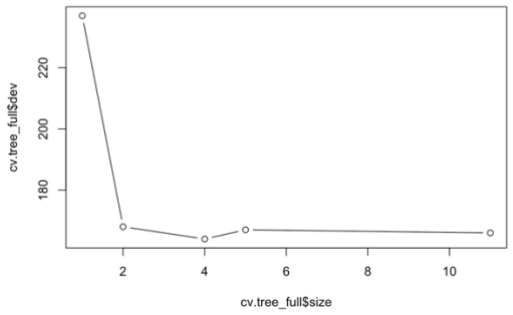
$\hat{y}_{R_m}$: predicted outcome (mean of training outcomes) for observations that fall into R_m

α: tree's **tuning** parameter, it controls the trade-off between complexity and quality of the fit to training data. The higher the α, the larger the penalty for having a complex tree.

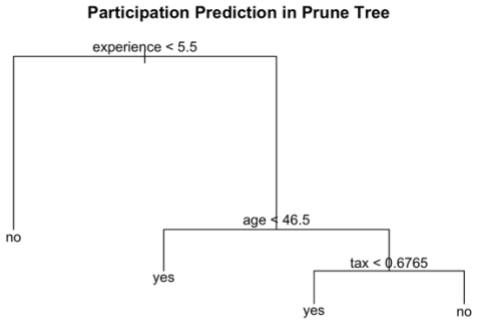
We can find a best T for each value of α

We need to find the best value of α. Again, we use cross validation to prune our tree: We perform cross-validation in order to determine the optimal level of tree complexity; cost complexity pruning is used in order to select a sequence of trees for consideration.

We can plot the error rate as a function of size:



We see from this plot that the tree with 4 terminal nodes results in the lowest cross-validation error rate, with 164 cross-validation errors. And we get the prune decision tree as follows:



If the experience is less than 5.5 years, the participation results would be “no”. For those who have experience larger than 5.5 years, if age is less than 46.5, the participation results would be “yes”. For those who have experience larger than 5.5 years and the age is larger than 46.5 and tax is larger than 0.6765, the participation results would be “no”.

This tree we have a height of 4 and accuracy rate of 0.662.

Compared with full tree, the prune tree will also decrease the error rate from the full tree. Hence, the tree models dose not perform well in this dataset.

In general, the decision trees are easy to explain, follow human decision-making more closely in some cases and can be displayed graphically can handle qualitative variables easier than regression. However, the decision trees do not have the level of prediction accuracy as other ML methods and tend to be very sensitive to the choice of sample.

7. Bagging

7.1 Bootstrap

Before introducing Bagging, we need to explain the definition of Bootstrap first.
we can use bootstrap to estimate the standard errors of our coefficients when they are not easily available as in coefficients of linear regressions where you get an automated output of SE in R. Bootstrap means sample your dataset with replacement several times.

7.2 Bagging

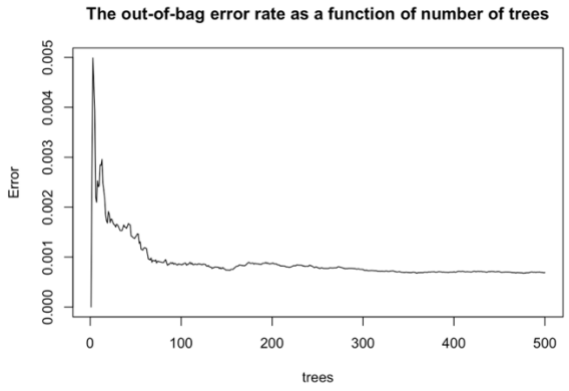
Bagging is short for Bootstrap Aggregation. We use Bootstrap to find the average outcome of multiple trees. Because normal decision trees suffer from having high variance (the prediction changes a lot with a small change in the sample). We can reduce this variance by averaging over multiple Bootstrap trees.

The objective function of bagging are as follows:

$$\hat{f}_{bag} = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(X)$$

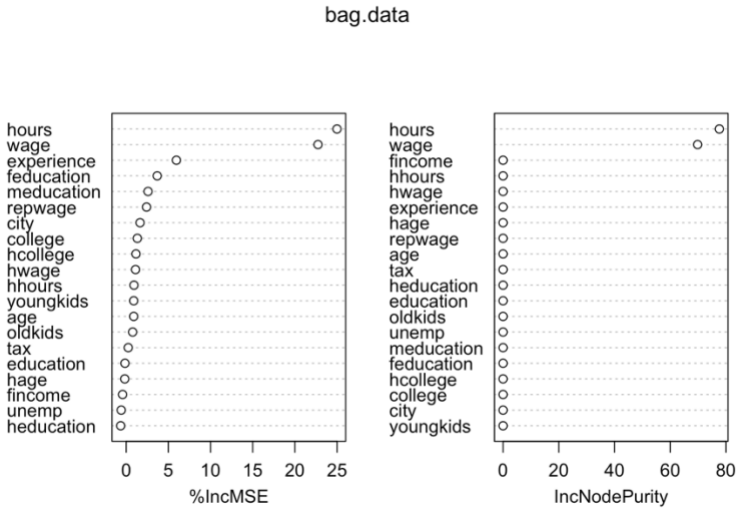
We use Bootstrap sample to fit each tree. On average each Bootstrap sample only includes around 2/3 of the data (check Exercise 2 of chapter 5). The remaining 1/3 are out-of-bag

(OOB) observations for the tree. We can plot the out-of-bag error rate as a function of number of trees:



In the plot, we can see the out-of-bag error rate as a function of number of trees. The error rate will largely increase at the beginning and largely decrease. And as the number of trees is more than around 70, the error rate is almost the same and lower than 0.001. The accuracy rate of this bagging model is 100%.

We can also plot the importance matrix of this model:



Two measures of variable importance are reported. The former is based upon the mean decrease of accuracy in predictions on the out-of-bag samples when a given variable is excluded from the model. The latter is a measure of the total increase in node impurity that results from splits over that variable, averaged over all trees. If we drop the variable, the mean decrease of accuracy in predictions on the out-of-bag samples will be 2.378246%. If we drop the variable, the mean decrease of accuracy in predictions on the out-of-bag samples will be 2.260392%. After visualizing the importance matrix, the results indicate that across all of the trees considered in the bagging classification, the hours and the wages are by far the two most important variables.

8. Random forest

8.1 Definition of Random Forest

There is one disadvantage of bagging. Suppose there is a very strong predictor in the model along with some moderate predictors. Each tree considers this strong predictor on top of its splits. The collection of generated trees at the end will look quite similar to each other. Averaging over correlated trees cannot reduce the variance very much.

Random Forest is an improvement over Bagging, they are very similar but with a very small but important difference. The difference is that each time that we split the tree, we select a random subset of predictors instead of all predictors.

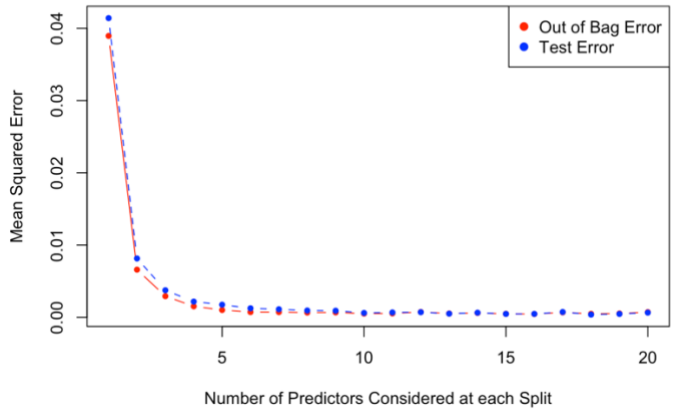
8.2 Hyperparameter tuning by looping over mtry

"Mtry" is the number of variables used at each split.

In practice, the split allows $m \approx \sqrt{p}$ or $m \approx p/3$. Now, instead of using random number for the "mtry" around square root of the total number of variables, we need to find the best number of

variables to use by estimating the Out of Bag (OOB) error. We use hyperparameter tuning by looping over “mtry” and comparing the error rate for each value.

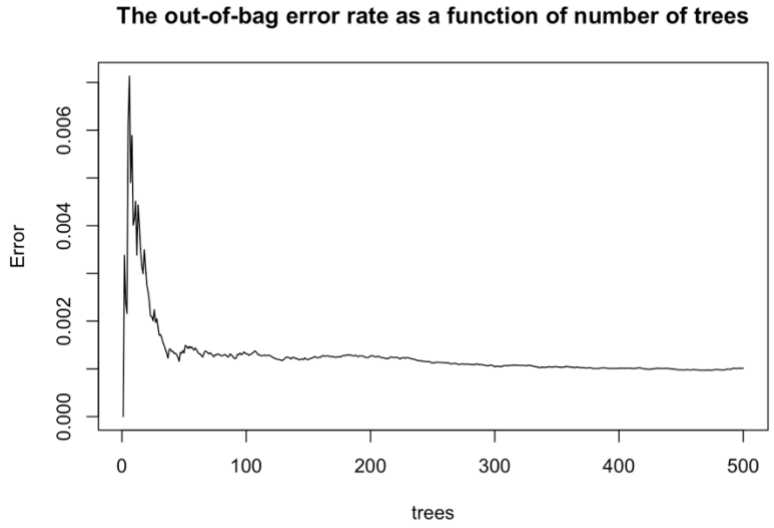
We can plot the test error and out-of-bag error in a same graph vs mtry:



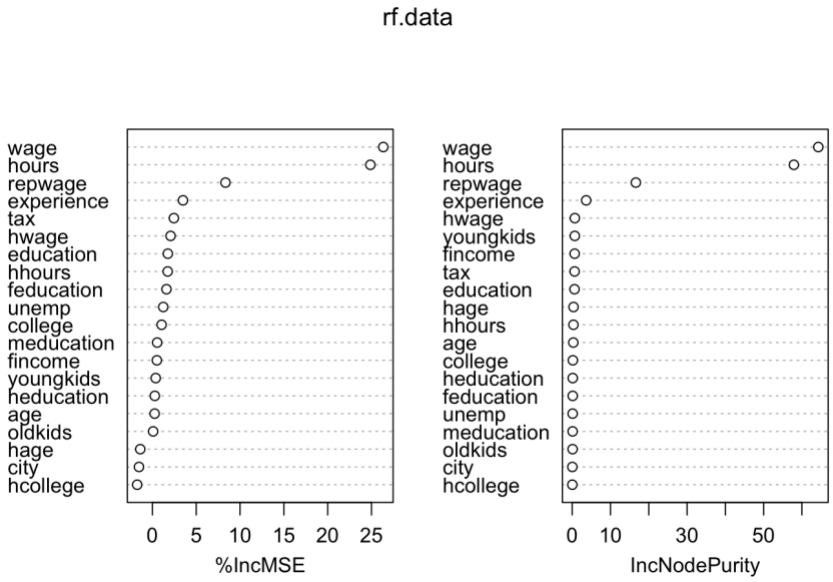
And we get mtry = 3 would be the best mtry.

8.3 Performance of the random forest

We can plot the out-of-bag error rate as a function of number of trees in random forest:



In this plot of the error rate as a function of number of trees, we can see that the error rate decreases gradually and after the number of trees go after 150, and the error rate stays at about 0.002. After running the random forest model, we get accuracy rate of 100%.



After visualizing the importance matrix, the results indicate that across all of the trees considered in the random forest classification, the hours and the wages are still by far the two most important variables, and we can see the repwage is the third most important variable.

9 Boosting classification

In boosting, the trees are trained sequentially based on the reside of the previous step. They grow sequentially in the sense that each tree uses the information from the previous tree. Unlike random forests and bagging, boosting can overfit if the number of trees (B) is very large. An appropriate choice of B can be made using cross-validation. The number of splits d controls the tree complexity.

The algorithm for boosting is shown below:

Algorithm 8.2 *Boosting for Regression Trees*

1. Set $\hat{f}(x) = 0$ and $r_i = y_i$ for all i in the training set.
2. For $b = 1, 2, \ldots, B$, repeat:
   (a) Fit a tree $\hat{f}^b$ with d splits ($d+1$ terminal nodes) to the training data (X, r).
   (b) Update $\hat{f}$ by adding in a shrunken version of the new tree:
   $$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x). \quad (8.10)$$
   (c) Update the residuals,
   $$r_i \leftarrow r_i - \lambda \hat{f}^b(x_i). \quad (8.11)$$
3. Output the boosted model,
   $$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x). \quad (8.12)$$

By running the boosting model, we get the accuracy rate of 100%.

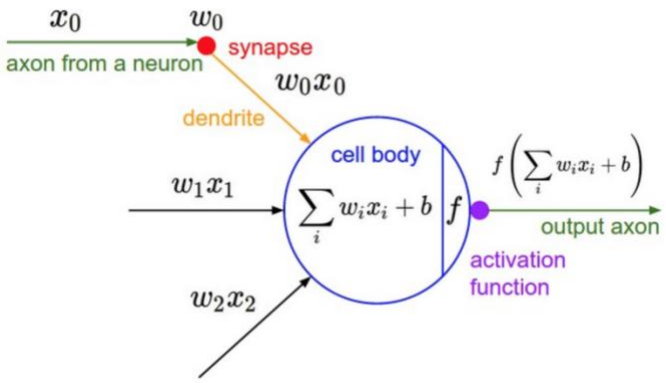
10. XGboost classification

XGBoost, or eXtreme Gradient Boosting, implements gradient boosting, but now includes a regularization parameter and implements parallel processing. It also has a built-in routine to handle missing values. XGBoost also allows one to use the model trained on the last iteration and updates it when new data becomes available.

The accuracy rate of XGboost classification is 100%.

11. Neural Net

The data goes through a network layer by layer until it produces an output. Each Neuron produces an output or activation. We want to find the right weights to connect Neurons. This is called fitting or training the model. The objective function of Neural Net is as follows:



Different functions of each layer and activity function can be used, such as ReLU, sigmoid, softmax.

And here we choose a model as follows: fully-connected (64 units), ReLU, fully-connected (64 units), ReLU, fully-connected (64 units), ReLU, fully-connected (2 units), softmax.

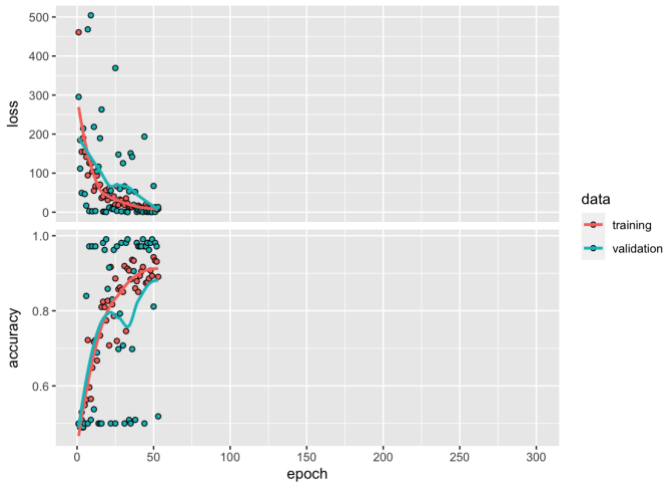
The objective function for ReLU is

$$\textbf{ReLU} \text{ function: } \sigma(v) = max(0, v)$$

And the objective function for softmax is

$$g_k(T) = \frac{e^{T_k}}{1 + e^{T_k}}$$

We can plot the training process of our model:



From this plot, we can see that the accuracy rate of Neural Net will increases after epochs and finally get to around 96%.

12. Comparing All Models

Models	Accurancy rate
Logistic(education+age+tax+experience+youngkids)	0.7317397;
Probit Classification	0.7343958
Lasso	0.9377778
Ridge	0.9111111
FULL Decision tree	0.6888889
Prune Tree	0.6622222
Boot-strap	0.9543805
Bagging classification	1
Random Forest classification	1
Boosting classification	1
XGboost classification	1
Boosting classification after tuning	1
Neural Net	0.96
KNN	0.7869917

From the chart above, we can see that Bagging classification, Random Forest classification, boosting classification, and XGboost classification will have the highest accuracy rate for 100% and neural net, lasso, ridge, and bootstrap also have high accuracy rate around 95%. And the decision trees, logit and probit do not perform well in this dataset.

Conclusion

After analyzing the dataset of labor force participation in 1976 by applying the machine learning methods, our analysis suggests that experience, tax, and age, presents of young kids will mostly influence the labor force participation.

Our research results from this dataset agree with the hypothesis of other papers in this field. For instance, the tax is highly correlated to policy, which proved that what Jacobs (2015) suggests: Policy, economic, and fiscal factors also will influence the labor force participation. Also, age and experience are the demographic factors mentioned by the Congressional Budget Office (2018). And the presence of young kids also largely affects the labor force participation, which agrees with Darian's theories.

We find that different models have different prediction performance in this model. For instance, Firstly, in terms of the correlation between each variable and participation, we can see that education and experience will positively affect the participation, and tax, number of young kids, will negatively affect the participation. Secondly, from the Logistic Classification, we can see that all the variables will affect the participation, so the full model with more variables would work the best and have the highest accuracy rate. And the experience would be the most determinant variable. According to the p-value of each variable, education, age, experience, youngkids have p-values lower than 0.001, so they are significantly different from zero, with a

confidence level larger than 99%. And tax would have the highest standard error. Thirdly, the accuracy of Probit regression and Logistic regression is very close. Probit regression's accuracy rate is slightly higher than the Logistic regression's accuracy rate, both of them are around 73%. Fourthly, for the KNN model, with K = 25, we have the highest accuracy rate, around 78%. Fifthly, for ridge and Lasso, we can see that the accuracy rate of the ridge model is 0.9309429, and the accuracy rate of the lasso model is 0.9163347. Both of them largely increase the accuracy rate of the models above. Sixthly, the decision tree, 2e see from this plot that the tree with 4 terminal nodes results in the lowest cross-validation error rate, with 164 cross-validation errors. The full tree will also increase the error rate from the full tree, with an accuracy rate of around 68%. Seventhly, the error rate of the Boosting classification, Bagging classification, XGboost classification, and Random Forest classification are the smallest, both are 0. And all the models have very low error rates. And Boosting classification, Bagging classification, XGboost classification, and Random Forest classification perform the best. Lastly, in the Neural Net model we use, the accuracy rate is around 96%. Although this might not be the best Neural Net that fits our dataset, the accuracy rate is fairly high.

In the future research, we would like to explore the different factors influence different genders in labor force participation. For instance, how would the presents of young kids influence the differences of labor force participants for woman and men. Another factor we want to explore is the marriage. How would marriage affect the labor force participant of people. Also, in the future, I would like to explore more about the factors of income in people's labor force participation.

References

Congressional Budget Office. (2018). Factors affecting the labor force participation of people ages 25 to 54. *Nonpartisan Analysis for the US Congress.* Retrieved from https://www.cbo.gov/publication/53452

Darian, J. (1976). Factors influencing the rising labor force participation rates of married women with pre-school children. *Social Science Quarterly*. Vol. 56, No. 4 (MARCH, 1976), pp. 614-630

Jacobs, E. (2015). The declining labor force participation rate: causes, consequences, and the path forward. *Washington Center for Equitable Growth.* Retrieved from https://equitablegrowth.org/declining-labor-force-participation-rate-causes-consequences-path-forward/

Greene, W.H. (2003). Econometric Analysis, 5th edition. Upper Saddle River, NJ: Prentice Hall.

McCullough, B.D. (2004). Some Details of Nonlinear Estimation. In: Altman, M., Gill, J., and McDonald, M.P.: *Numerical Issues in Statistical Computing for the Social Scientist*. Hoboken, NJ:John Wiley, Ch. 8, 199–218.

Mroz, T.A. (1987). The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions. *Econometrica*, 55, 765–799.

Winkelmann, R., and Boes, S. (2009). Analysis of Microdata, 2nd ed. *Berlin and Heidelberg*: Springer-Verlag.

Wooldridge, J.M. (2002). Econometric Analysis of Cross-Section and Panel Data. Cambridge, MA: MIT Press.