A STUDY ON WAGES BASED ON EXPERIENCE, EDUCATION, GENDER, AND TENURE AT THE COMPANY

Conducted by Sarang Deshpande

University of North Carolina at Charlotte
STAT 2223-002

Introduction

For this project, I wanted to explore variable selection through all possible regression, backward elimination, and forward selection. These are various selection techniques that are used to determine which variables are statistically significant enough to be considered in the model. The procedure for each selection method slightly varies, however. For backwards selection, all variables are included in the model and are then tested using their t-statistic or F-statistic. The variable with the lowest F/T-statistic is then removed until the model reaches a satisfactory level of accuracy. For forward selection, the variables are added in based on their t-score; if their t-score is significant, they are added to the model. The all-possible regression technique uses a certain criterion (R^2, c_p, etc.) to determine if the variable is significant to an accurate model.

To explore the concept of variable selection, I used data from the Wooldridge package that is within R. I specifically decided to use the wage1 dataset, which explored how wages are affected by different factors such as tenure at the company, education, gender, and job experience. In order to call the dataset, I had to use the following code:

```
install.packages("wooldridge")
library(wooldridge)
data("wage1", package="wooldridge")
force(wage1)
```

Because the dataset is in a package, I had to use the force function to open the dataset and be able to view it.

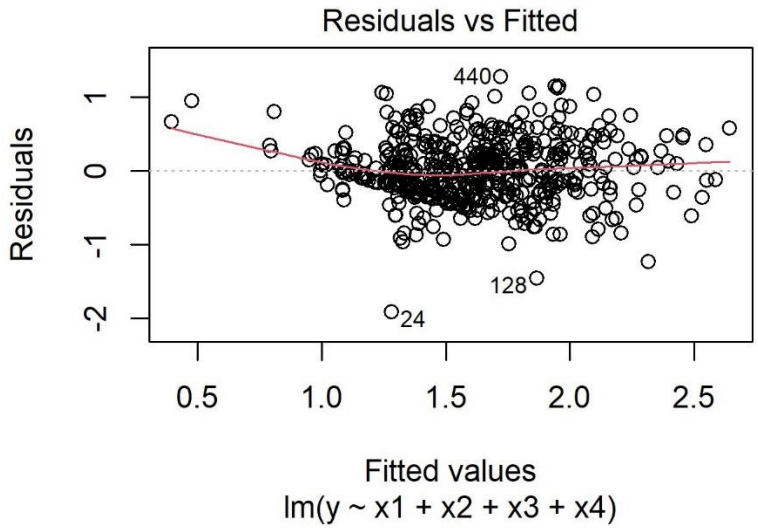
Review of Methodology

After I imported the package and dataset into R, I first decided to assign variables to make it easier to craft the various linear models for the project. Once I assigned the variables, I created a linear model representing the relationship between wages and the various independent

variables (Experience, Education, Gender, and Tenure at the Company). I then found the coefficients and the summary statistics of the linear model I generated. I also decided to plot the data to better visualize the data; upon seeing the plot, I noticed that the data had a second-order model. In order to get a first-order model, I decided to set the dependent variable as the logarithm of the wages instead of just wages. This made it simpler to formulate the beta coefficients and decreased the distance of certain residuals.

Once I had created the linear model, I imported the leaps library in order to access the variable screening functions, such as regsubset(). Once I had imported the leaps library, I then performed the three variable selection techniques mentioned above: all possible regression, forward selection, and backward elimination. I stored each of the selection techniques in an object and created summaries for each of the techniques as well. The final part of my methodology was to store the R^2, c_p, and BIC value into a data frame. These are the criterion used in the all-possible regression method.

Data Analysis



The image above is the plot of the regression against the 526 observations in the dataset. From the coefficient function, I determined that the regression equation is:

$$\ln(wages) = 0.501347968 + 0.087462324(Education) + 0.004629381(Experience) + 0.017366965(Tenure) - 0.301145873(Gender)$$

From the various variable selection methods, I determined that all four of the variables in the model above are statistically significant to determine wages, and thus removed no variables from the model.

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.91588 -0.26192 -0.02399  0.26349  1.27352

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.501348   0.101902   4.920 1.16e-06 ***
x1           0.087462   0.006939  12.605  < 2e-16 ***
x2           0.004629   0.001627   2.845  0.00461 **
x3           0.017367   0.002976   5.835 9.45e-09 ***
x4          -0.301146   0.037246  -8.085 4.37e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.416 on 521 degrees of freedom
Multiple R-squared:  0.3923,    Adjusted R-squared:  0.3876
F-statistic: 84.07 on 4 and 521 DF,  p-value: < 2.2e-16
```

The above screengrab highlights other useful information about the model. We can see that this model has an R^2 of 39.23% and a standard deviation of 0.416.

Concluding Remarks

Overall, this project delved into the concept of variable selection. To ensure thorough understanding, I performed three different methods: all-possible regression, backward

elimination, and forward selection. The data I selected was econometric data in the Wooldridge package, specifically the wage1 dataset. From this process, I determined that all four of the independent variables I selected were statistically significant and were important to the model. In addition, I was able to learn more about the different criterion used for the all-possible regression, such as c_p, R^2, or BIC.

Code:

```
#Setting up data from the wooldridge package. Wooldridge is an econometric data
#package. I installed the package, opened the library, and extracted the "wage1"
#dataset.

install.packages("wooldridge")
library(wooldridge)
data("wage1", package="wooldridge")
force(wage1)
mydata<-wage1

#Creating variables and a linear model based on the variables
y<-wage1$lwage
x1<-wage1$educ
x2<-wage1$exper
x3<-wage1$tenure
x4<-wage1$female
lwage_estimation<-lm(y~x1+x2+x3+x4, data=wage1)
plot(lwage_estimation)

#Finding the regression coefficients and other summary/anova statistics
coefficients(lwage_estimation)
summary(lwage_estimation)
anova(lwage_estimation)

#Using all possible regression, backwards elimination, and forward selection
#methods to determine what variables are useful for the model
library(leaps)
```

```
#All Possible Regression
apr<-regsubsets(y~x1+x2+x3+x4,
            data=mydata,
            nvmax=3,
            intercept=TRUE,
            method="exhaustive")
apr
apr.sum<-summary(apr)
apr.sum

#Backwards
bcw<-regsubsets(y~x1+x2+x3+x4,
            data=mydata,
            nvmax=3,
            intercept=TRUE,
            method="backward")
bcw
bcw.sum<-summary(bcw)
bcw.sum

#Forwards
frw<-regsubsets(y~x1+x2+x3+x4,
            data=mydata,
            nvmax=3,
            intercept=TRUE,
            method="forward")
frw
frw.sum<-summary(frw)
frw.sum

#Other information for hypothesis testing. For simplicity, only taking from APR
data.frame(
  Adj.R2 = which.max(apr.sum$adjr2),   # Adjusted R-squres
  CP = which.min(apr.sum$cp),          # Cp
  BIC = which.min(apr.sum$bic)         # BIC
)
```