**Statistical Modeling for Business Analytics – MBA652A – Project 3**

# BINARY DEPENDENT VARIABLE DATA: Factors affecting women employment

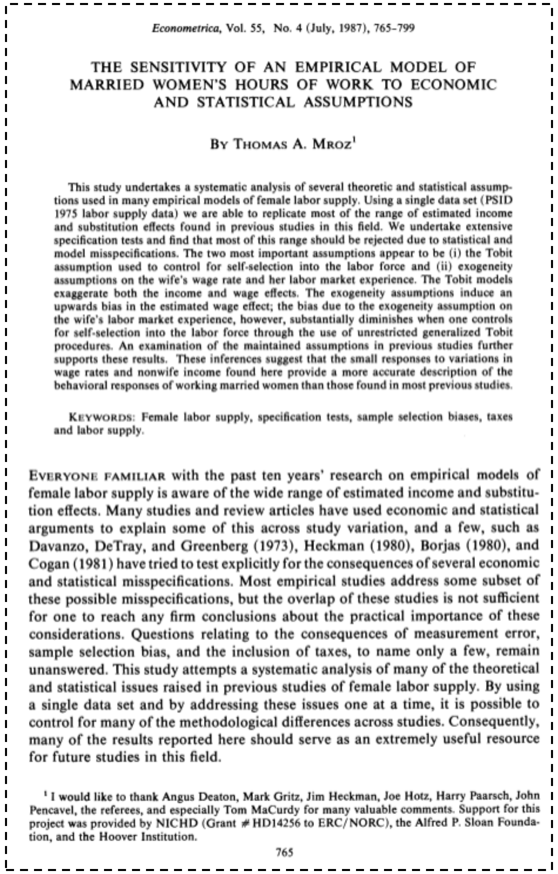
**Submitted To:**
**Prof. (Dr.) Devlina Chatterjee**


**Submitted By: Group 5**

1. Ashish Tiwari (21129004)
2. Jyoti Sharma (21129265)
3. Shiv Shakti Singh (21129024)
4. Shreeyash Nitin Malode (20214271)

# Outline of the Presentation

1. Introduction
2. Objective
3. Descriptive Analysis
4. Scatter plot & Correlation Matrix
5. Models
6. Interpretation of Results
7. Inference & Conclusion



*Econometrica*, Vol. 55, No. 4 (July, 1987), 765–799

THE SENSITIVITY OF AN EMPIRICAL MODEL OF MARRIED WOMEN'S HOURS OF WORK TO ECONOMIC AND STATISTICAL ASSUMPTIONS

BY THOMAS A. MROZ[1]

This study undertakes a systematic analysis of several theoretic and statistical assumptions used in many empirical models of female labor supply. Using a single data set (PSID 1975 labor supply data) we are able to replicate most of the range of estimated income and substitution effects found in previous studies in this field. We undertake extensive specification tests and find that most of this range should be rejected due to statistical and model misspecifications. The two most important assumptions appear to be (i) the Tobit assumption used to control for self-selection into the labor force and (ii) exogeneity assumptions on the wife's wage rate and her labor market experience. The Tobit models exaggerate both the income and wage effects. The exogeneity assumptions induce an upwards bias in the estimated wage effect; the bias due to the exogeneity assumption on the wife's labor market experience, however, substantially diminishes when one controls for self-selection into the labor force through the use of unrestricted generalized Tobit procedures. An examination of the maintained assumptions in previous studies further supports these results. These inferences suggest that the small responses to variations in wage rates and nonwife income found here provide a more accurate description of the behavioral responses of working married women than those found in most previous studies.

KEYWORDS: Female labor supply, specification tests, sample selection biases, taxes and labor supply.

EVERYONE FAMILIAR with the past ten years' research on empirical models of female labor supply is aware of the wide range of estimated income and substitution effects. Many studies and review articles have used economic and statistical arguments to explain some of this across study variation, and a few, such as Davanzo, DeTray, and Greenberg (1973), Heckman (1980), Borjas (1980), and Cogan (1981) have tried to test explicitly for the consequences of several economic and statistical misspecifications. Most empirical studies address some subset of these possible misspecifications, but the overlap of these studies is not sufficient for one to reach any firm conclusions about the practical importance of these considerations. Questions relating to the consequences of measurement error, sample selection bias, and the inclusion of taxes, to name only a few, remain unanswered. This study attempts a systematic analysis of many of the theoretical and statistical issues raised in previous studies of female labor supply. By using a single data set and by addressing these issues one at a time, it is possible to control for many of the methodological differences across studies. Consequently, many of the results reported here should serve as an extremely useful resource for future studies in this field.

[1] I would like to thank Angus Deaton, Mark Gritz, Jim Heckman, Joe Hotz, Harry Paarsch, John Pencavel, the referees, and especially Tom MaCurdy for many valuable comments. Support for this project was provided by NICHD (Grant #HD14256 to ERC/NORC), the Alfred P. Sloan Foundation, and the Hoover Institution.



**Main Reference -** 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.

**Secondary Reference -**

https://sites.google.com/site/econometricsacademy/masters-econometrics/probit-and-logit-models?authuser=0

**Dataset Source**
https://vincentarelbundock.github.io/Rdatasets/datasets.html

**Software used -** R & Excel

**Image Source** – Screenshot of & dataset website - https://hrsonline.isr.umich.edu

# Introduction

- There are many direct and indirect factor which can affect employment of a women. Direct factors like her age, education, wage rate and number of toddlers etc. and indirect factors like her parent's education, her family income, her husband age and wages etc.
- Dataset is collection of survey by Dr. Thomas Mroz for his research paper.
- It has a sample of 753 married white women between the ages of 30 and 60 in 1975.
- In 1975, a total of 428 women were working at some time during the year
- Total number of observations : 753

```
> table(MROZ$inlf)

  0   1
325 428
```

Checking whether data has gap or not ? **No gap in data**

- Variables under focus :
- **inlf –** employed=1 or unemployment=0
- **kidslt6 –** number of kids less than 6 years of age in household
- **Age -** women's age
- **educ –** women years of education educational attainment, in years
- **nwifeinc -** non-wife family income, in 1975 dollars; family income excluding women's income
- **exper -** previous work experience

**Image Source** – Computed R Output

# Objective

- The objective is to investigate influence of direct and indirect factors like family income, age, education years, previous work experience and age of children on women employment using probit and logit model.
- **Null Hypothesis ($H_0$) :** No relationship exist between women employment (inlf), and family income, age, education years, previous work experience and age of children of a women.
- **Alternate Hypothesis ($H_1$)** : Relationship exist between women employment (inlf) , and family income, age, education years, previous work experience and age of children of a women.

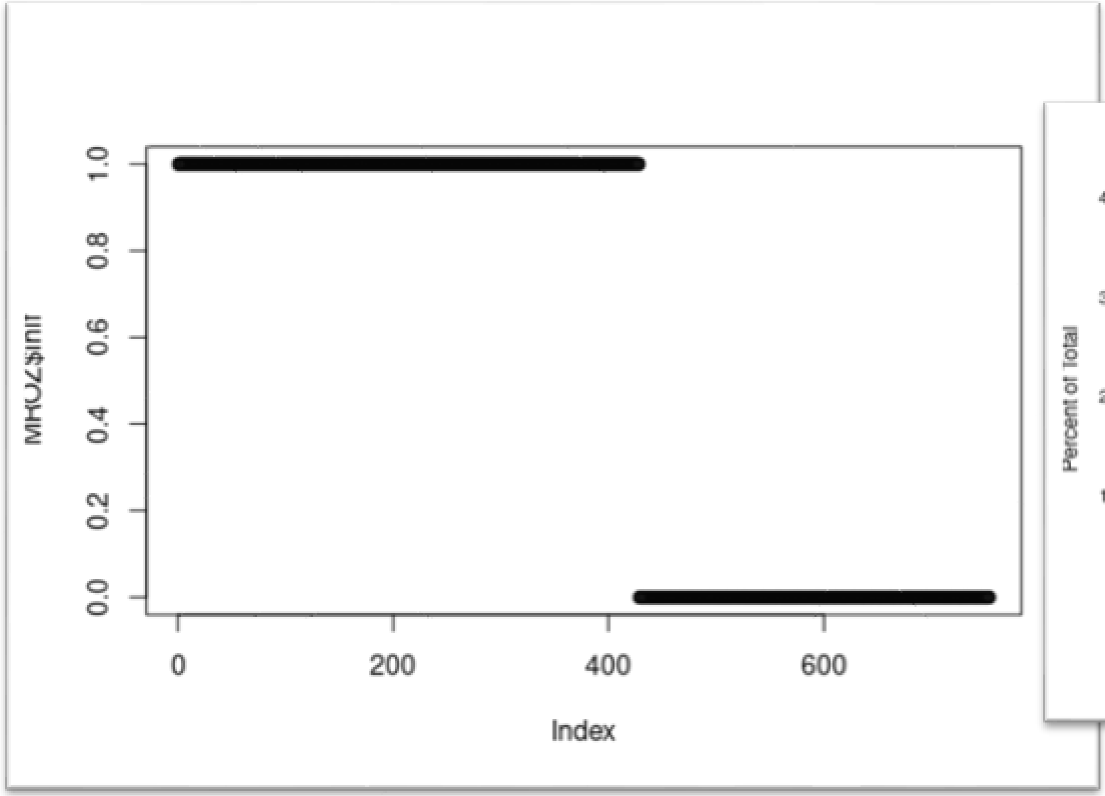
**Image Source** –

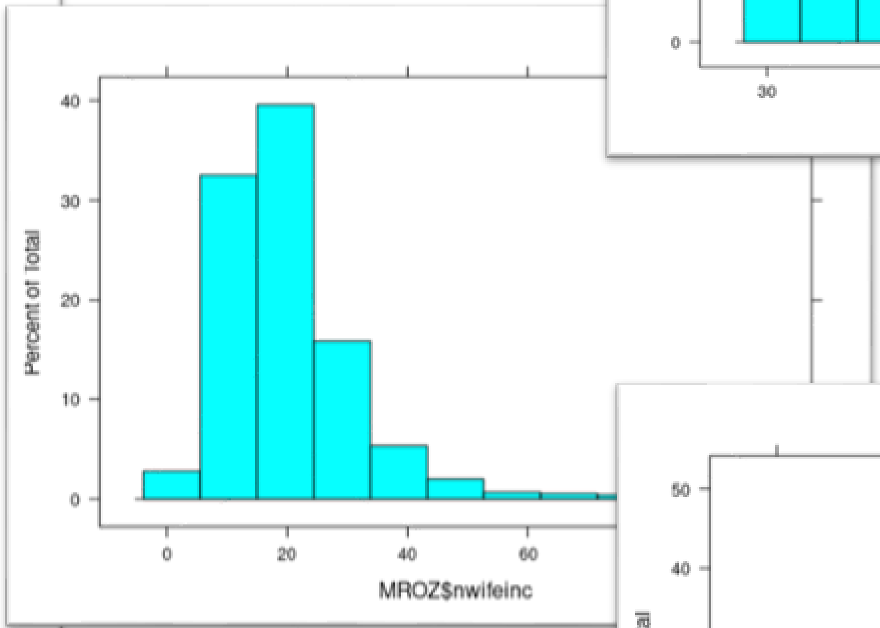
# Descriptive Analysis

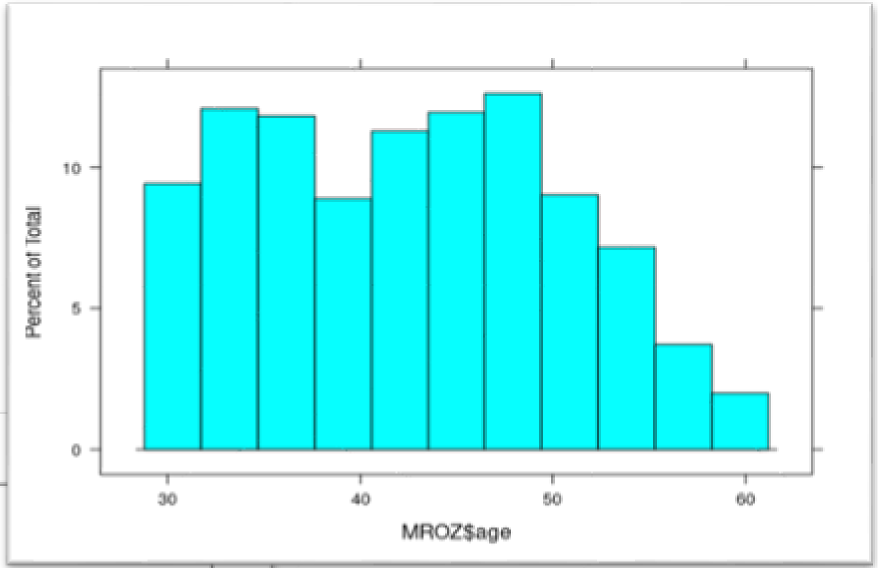
| | inlf | kidslt6 | age | educ | city | exper | nwifeinc |
|---|---|---|---|---|---|---|---|
| **Mean** | 0.57 | 0.24 | 42.54 | 12.29 | 0.64 | 10.63 | 20.13 |
| **Standard Error** | 0.02 | 0.02 | 0.29 | 0.08 | 0.02 | 0.29 | 0.42 |
| **Median** | 1.00 | 0.00 | 43.00 | 12.00 | 1.00 | 9.00 | 17.70 |
| **First Quartile** | 0.00 | 0.00 | 36.00 | 12.00 | 0.00 | 4.00 | 13.03 |
| **Third Quartile** | 1.00 | 0.00 | 49.00 | 13.00 | 1.00 | 15.00 | 24.47 |
| **Variance** | 0.25 | 0.27 | 65.17 | 5.20 | 0.23 | 65.11 | 135.37 |
| **Standard Deviation** | 0.50 | 0.52 | 8.07 | 2.28 | 0.48 | 8.07 | 11.63 |
| **Kurtosis** | -1.93 | 5.30 | -1.02 | 0.76 | -1.65 | 0.71 | 8.45 |
| **Skewness** | -0.28 | 2.31 | 0.15 | 0.02 | -0.60 | 0.96 | 2.21 |
| **Range** | 1.00 | 3.00 | 30.00 | 12.00 | 1.00 | 45.00 | 96.03 |
| **Minimum** | 0.00 | 0.00 | 30.00 | 5.00 | 0.00 | 0.00 | -0.03 |
| **Maximum** | 1.00 | 3.00 | 60.00 | 17.00 | 1.00 | 45.00 | 96.00 |
| **Sum** | 428.00 | 179.00 | 32031.00 | 9252.00 | 484.00 | 8005.00 | 15157.11 |

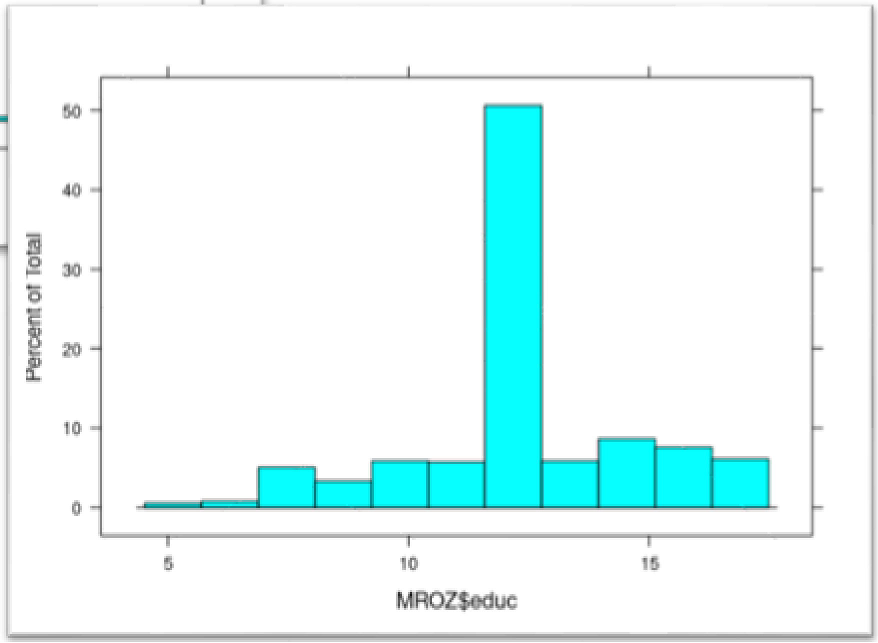
**Source** – Computed R & Excel Output

# Descriptive Analysis (cond.)



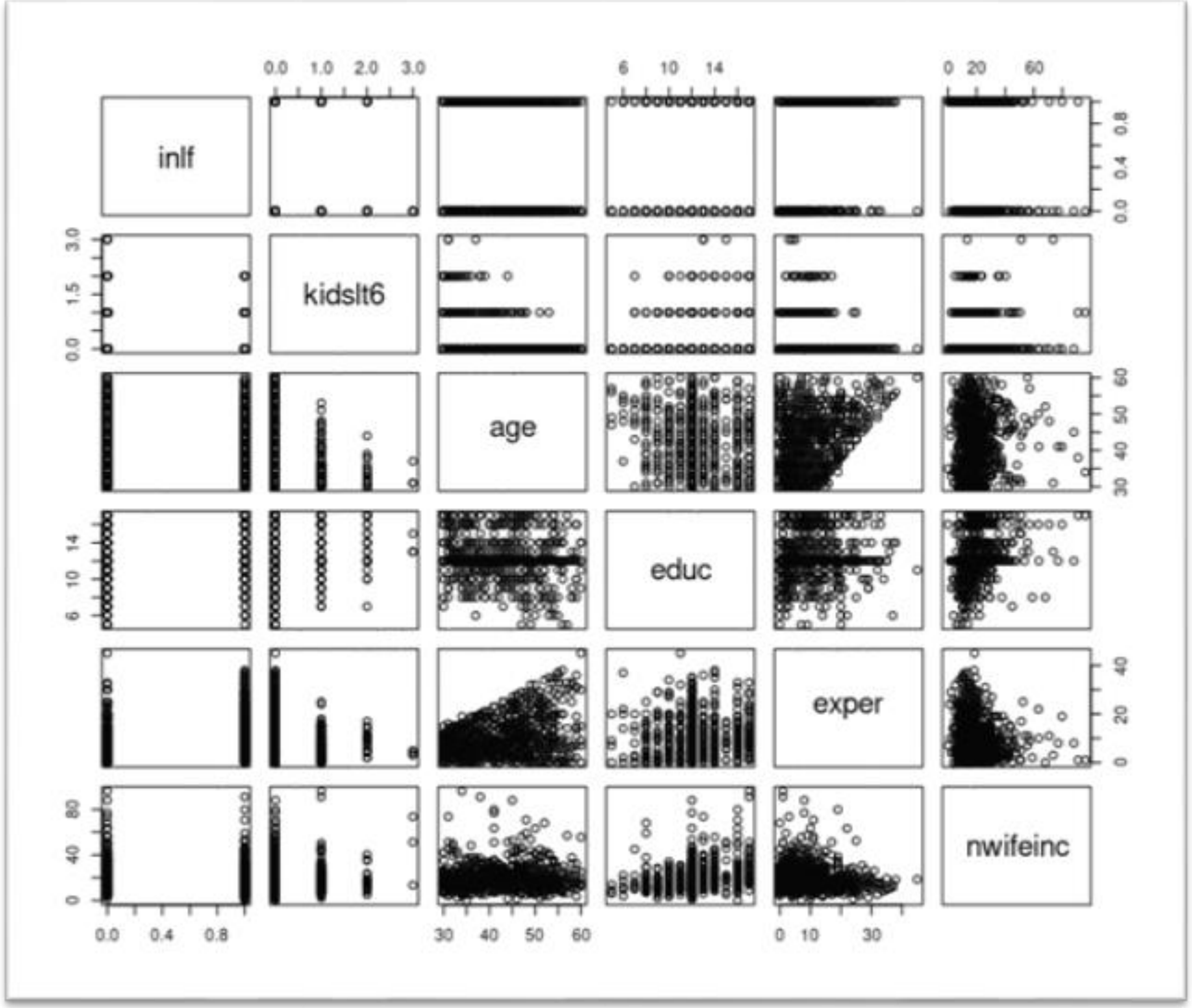



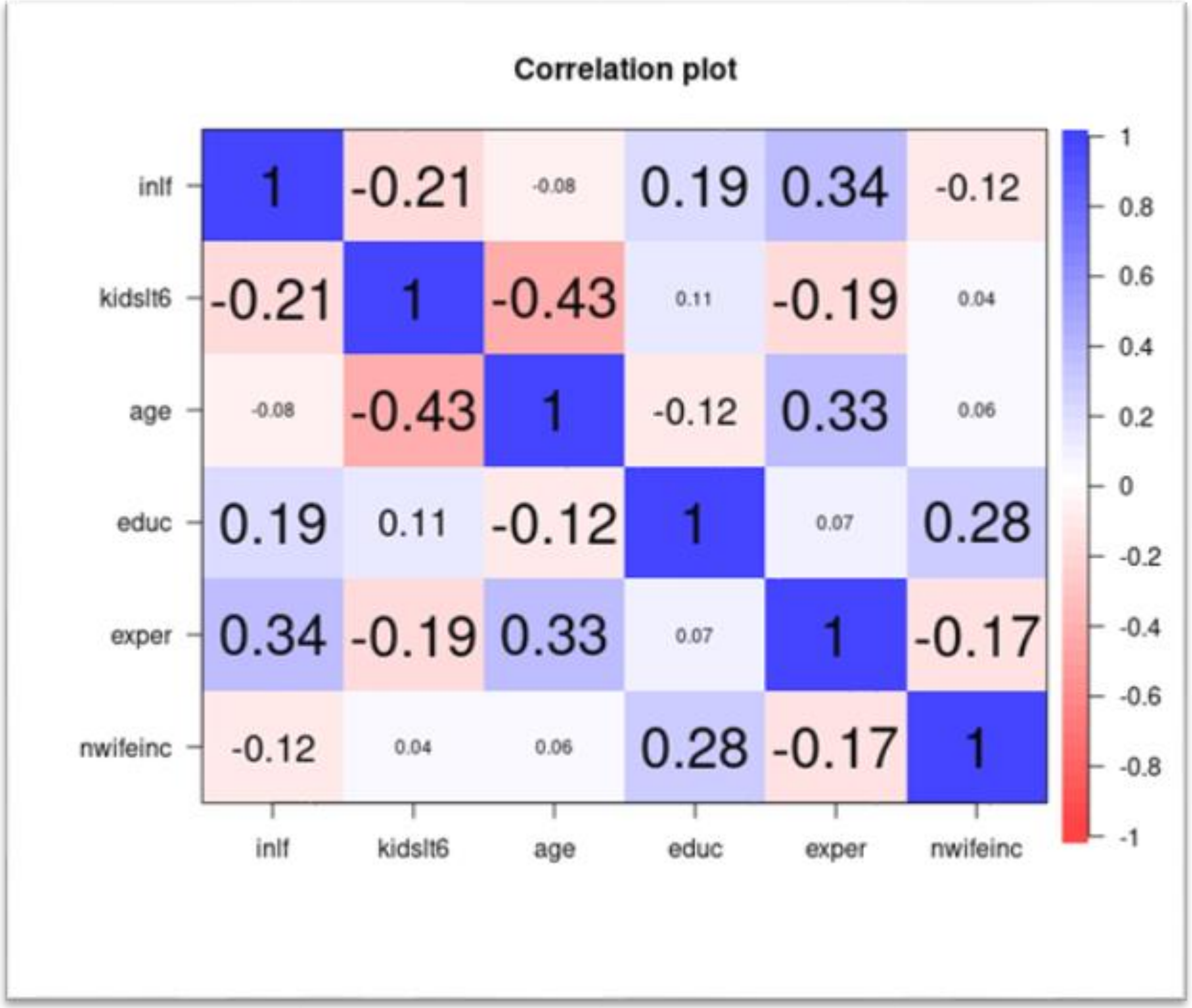




**Source** – Computed R & Excel Output

# Scatter plot & correlation matrix




**Source** – Computed R & Excel Output

# Linear Probability Model (OLS Regression)

```
> LPM <- lm(inlf ~ nwifeinc + educ + exper + age + kidslt6, MROZ)
> summary(LPM)

Call:
lm(formula = inlf ~ nwifeinc + educ + exper + age + kidslt6,
    data = MROZ)

Residuals:
     Min       1Q   Median       3Q      Max
-1.06854 -0.38575  0.08284  0.35737  0.95335

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.769833   0.135007   5.702 1.70e-08 ***
nwifeinc    -0.003259   0.001456  -2.239   0.0255 *
educ         0.039129   0.007364   5.314 1.42e-07 ***
exper        0.022211   0.002144  10.358  < 2e-16 ***
age         -0.018508   0.002299  -8.052 3.22e-15 ***
kidslt6     -0.275306   0.033366  -8.251 7.08e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4298 on 747 degrees of freedom
Multiple R-squared:  0.253,     Adjusted R-squared:  0.248
F-statistic: 50.61 on 5 and 747 DF,  p-value: < 2.2e-16
```

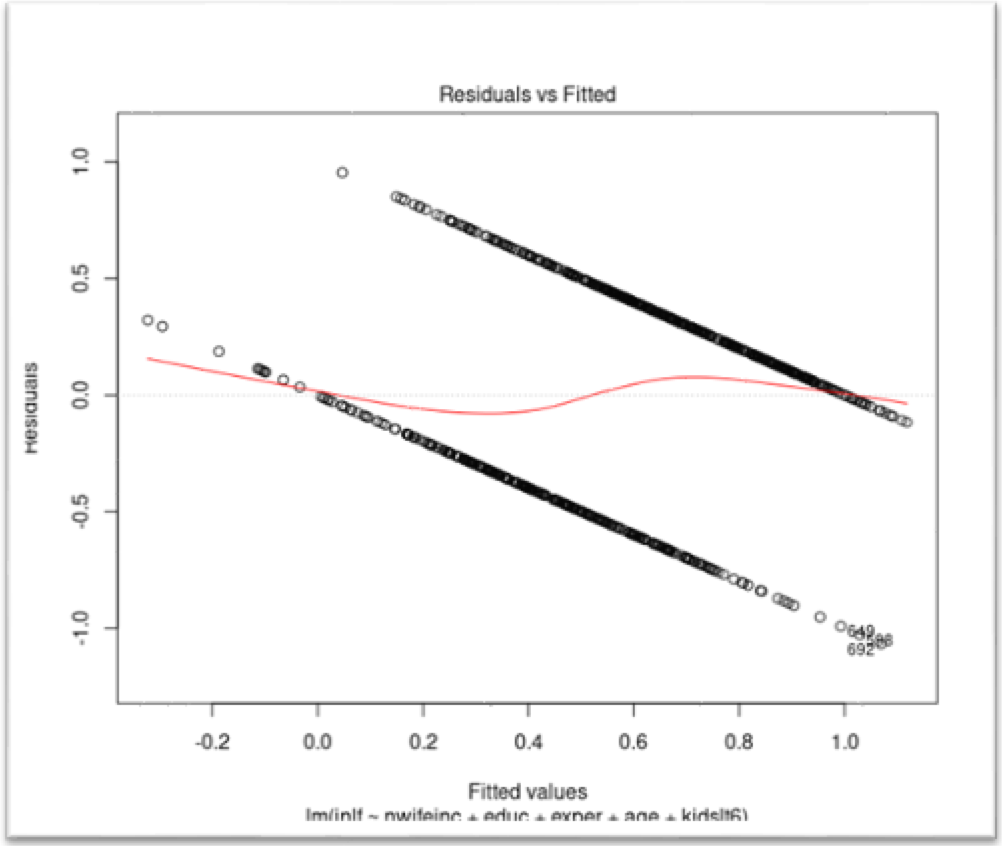



**Image Source** – Computed R Output

# Logit model 1 & 2

```
Call:
glm(formula = inlf ~ nwifeinc + exper + kidslt6, family = binomial(link = "logit"),
    data = MROZ)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7738  -1.0533   0.5607   1.0008   2.0605

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.255988   0.214133  -1.195   0.2319
nwifeinc    -0.012919   0.007098  -1.820   0.0687 .
exper        0.096011   0.012214   7.861 3.81e-15 ***
kidslt6     -0.659761   0.159723  -4.131 3.62e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1029.7  on 752  degrees of freedom
Residual deviance:  909.8  on 749  degrees of freedom
AIC: 917.8

Number of Fisher Scoring iterations: 4
```

```
Call:
glm(formula = inlf ~ nwifeinc + educ + exper + kidslt6, family = binomial(link = "logit"),
    data = MROZ)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7290  -1.0426   0.5211   0.9585   2.2444

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.885372   0.490226  -5.886 3.96e-09 ***
nwifeinc    -0.029424   0.008096  -3.634 0.000279 ***
educ         0.250997   0.041498   6.048 1.46e-09 ***
exper        0.089432   0.012366   7.232 4.75e-13 ***
kidslt6     -0.854543   0.170025  -5.026 5.01e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1029.75  on 752  degrees of freedom
Residual deviance:  869.21  on 748  degrees of freedom
AIC: 879.21

Number of Fisher Scoring iterations: 4
```

**Image Source** – Computed R Output

# Logit model 3 (all variables)

```
> summary(Logit)

Call:
glm(formula = inlf ~ nwifeinc + educ + exper + age + kidslt6, 
    family = binomial(link = "logit"), data = MROZ)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.5175  -0.9174   0.4441   0.8841   2.2974  

Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept)  1.153219   0.742068   1.554   0.1202    
nwifeinc    -0.019900   0.008268  -2.407   0.0161 *  
educ         0.223366   0.042969   5.198 2.01e-07 ***
exper        0.117887   0.013386   8.807  < 2e-16 ***
age         -0.095141   0.013439  -7.080 1.45e-12 ***
kidslt6     -1.463577   0.200353  -7.305 2.77e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1029.75  on 752  degrees of freedom
Residual deviance:  812.92  on 747  degrees of freedom
AIC: 824.92

Number of Fisher Scoring iterations: 4
```

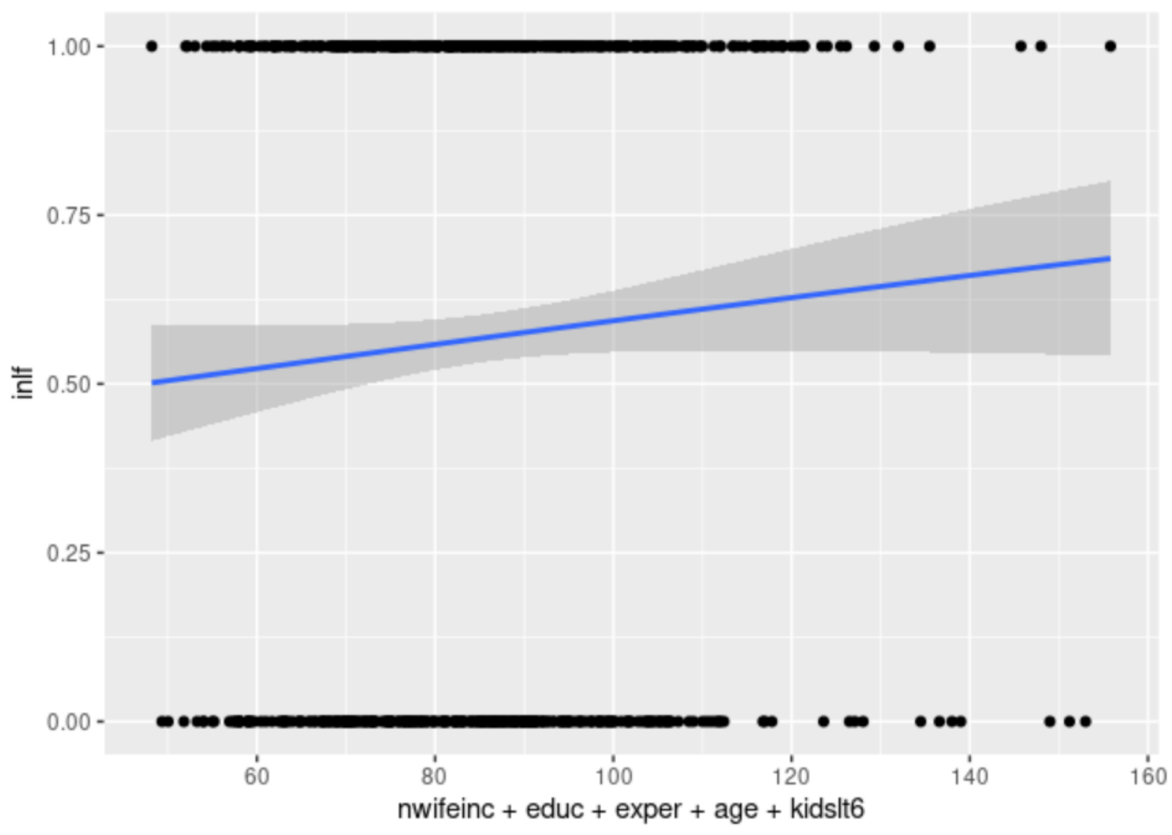



**Image Source** – Computed R Output

# Probit model 1 & 2

```
Call:
glm(formula = inlf ~ nwifeinc + exper + kidslt6, family = binomial(link = "probit"),
    data = MROZ)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9048  -1.0708   0.5806   1.0099   2.0629

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.128311   0.129962  -0.987   0.3235
nwifeinc    -0.007664   0.004272  -1.794   0.0728 .
exper        0.054396   0.006948   7.829 4.92e-15 ***
kidslt6     -0.408098   0.096101  -4.247 2.17e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1029.75  on 752  degrees of freedom
Residual deviance:  912.84  on 749  degrees of freedom
AIC: 920.84

Number of Fisher Scoring iterations: 3
```

```
Call:
glm(formula = inlf ~ nwifeinc + educ + exper + kidslt6, family = binomial(link = "probit"),
    data = MROZ)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8751  -1.0587   0.5294   0.9696   2.2631

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.717589   0.287755  -5.969 2.39e-09 ***
nwifeinc    -0.017238   0.004750  -3.629 0.000284 ***
educ         0.151236   0.024240   6.239 4.40e-10 ***
exper        0.050471   0.007056   7.152 8.53e-13 ***
kidslt6     -0.522148   0.100875  -5.176 2.26e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1029.75  on 752  degrees of freedom
Residual deviance:  871.59  on 748  degrees of freedom
AIC: 881.59

Number of Fisher Scoring iterations: 4
```

**Image Source** – Computed R Output

# Probit model 3 (all variables)

```
> summary(Probit)

Call:
glm(formula = inlf ~ nwifeinc + educ + exper + age + kidslt6,
    family = binomial(link = "probit"), data = MROZ)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5942  -0.9371   0.4342   0.8934   2.3229

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.764541   0.439490   1.740   0.0819 .
nwifeinc    -0.011371   0.004857  -2.341   0.0192 *
educ         0.131532   0.025082   5.244 1.57e-07 ***
exper        0.069148   0.007556   9.151  < 2e-16 ***
age         -0.057919   0.007790  -7.435 1.04e-13 ***
kidslt6     -0.886208   0.116696  -7.594 3.10e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1029.75  on 752  degrees of freedom
Residual deviance:  813.08  on 747  degrees of freedom
AIC: 825.08

Number of Fisher Scoring iterations: 4
```

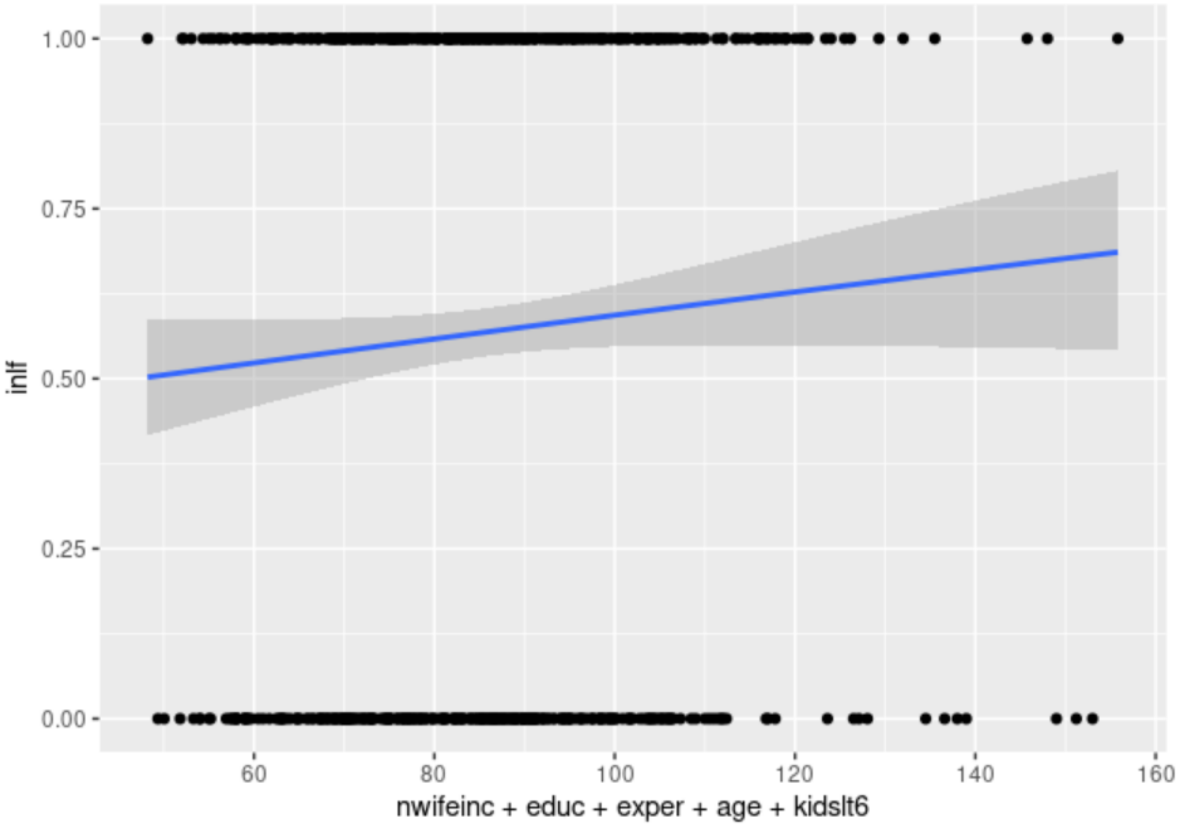



**Image Source** – Computed R Output

# Comparison of odds ratio

```
> exp(LPM$coefficients)%>% round(2)
(Intercept)    nwifeinc        educ       exper         age     kidslt6
       2.16        1.00        1.04        1.02        0.98        0.76
> exp(Logit$coefficients)%>% round(2)
(Intercept)    nwifeinc        educ       exper         age     kidslt6
       3.17        0.98        1.25        1.13        0.91        0.23
> exp(Probit$coefficients)%>% round(2)
(Intercept)    nwifeinc        educ       exper         age     kidslt6
       2.15        0.99        1.14        1.07        0.94        0.41
```

**Image Source** – Computed R Output

# Predicted probability & average marginal effect

```
> #Coefficients are marginal effects in a linear model
> coef(LPM)
 (Intercept)     nwifeinc         educ        exper          age      kidslt6 
 0.769832596 -0.003258636  0.039128569  0.022210504 -0.018507859 -0.275306267 
> summary(Probit.atmean)
   factor   inlf nwifeinc    educ   exper     age kidslt6     AME     SE       z      p   lower   upper
      age 0.5684  20.1290 12.2869 10.6308 42.5378  0.2377 -0.0226 0.0030 -7.4251 0.0000 -0.0286 -0.0166
     educ 0.5684  20.1290 12.2869 10.6308 42.5378  0.2377  0.0513 0.0098  5.2427 0.0000  0.0321  0.0705
    exper 0.5684  20.1290 12.2869 10.6308 42.5378  0.2377  0.0270 0.0029  9.1961 0.0000  0.0212  0.0327
  kidslt6 0.5684  20.1290 12.2869 10.6308 42.5378  0.2377 -0.3457 0.0457 -7.5682 0.0000 -0.4352 -0.2561
 nwifeinc 0.5684  20.1290 12.2869 10.6308 42.5378  0.2377 -0.0044 0.0019 -2.3407 0.0192 -0.0081 -0.0007
> summary(Logit.atmean)
   factor   inlf nwifeinc    educ   exper     age kidslt6     AME     SE       z      p   lower   upper
      age 0.5684  20.1290 12.2869 10.6308 42.5378  0.2377 -0.0231 0.0033 -7.0706 0.0000 -0.0294 -0.0167
     educ 0.5684  20.1290 12.2869 10.6308 42.5378  0.2377  0.0541 0.0104  5.1986 0.0000  0.0337  0.0745
    exper 0.5684  20.1290 12.2869 10.6308 42.5378  0.2377  0.0286 0.0032  8.9082 0.0000  0.0223  0.0348
  kidslt6 0.5684  20.1290 12.2869 10.6308 42.5378  0.2377 -0.3546 0.0488 -7.2723 0.0000 -0.4501 -0.2590
 nwifeinc 0.5684  20.1290 12.2869 10.6308 42.5378  0.2377 -0.0048 0.0020 -2.4066 0.0161 -0.0087 -0.0009
```

**Image Source** – Computed R Output

# Pseudo R-squared

**Logit**

```
> # Log-likelihood for model with only constant
> (LL0 <- logLik(r.model))
'log Lik.' -514.8732 (df=1)
> # Calculate pseudo R-squared
> (pseudo_r2 <- 1 - LLur/LL0)
'log Lik.' 0.2105663 (df=6)
```

**Probit**

```
> # Log-likelihood for model with only constant
> (LL0 <- logLik(r.model))
'log Lik.' -514.8732 (df=1)
> # Calculate pseudo R-squared
> (pseudo_r2 <- 1 - LLur/LL0)
'log Lik.' 0.2104035 (df=6)
```

**Image Source** – Computed R Output

# Confusion matrix logit model 1 & 3

```
> confusionMatrix(Logit1.pred, actual, positive = "1")
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 198  97
         1 127 331

               Accuracy : 0.7025
                 95% CI : (0.6685, 0.735)
    No Information Rate : 0.5684
    P-Value [Acc > NIR] : 2.417e-14

                  Kappa : 0.3869

 Mcnemar's Test P-Value : 0.05267

            Sensitivity : 0.7734
            Specificity : 0.6092
         Pos Pred Value : 0.7227
         Neg Pred Value : 0.6712
             Prevalence : 0.5684
         Detection Rate : 0.4396
   Detection Prevalence : 0.6082
      Balanced Accuracy : 0.6913

       'Positive' Class : 1
```

```
> confusionMatrix(Logit.pred, actual, positive = "1")
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 211  83
         1 114 345

               Accuracy : 0.7384
                 95% CI : (0.7054, 0.7695)
    No Information Rate : 0.5684
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.4606

 Mcnemar's Test P-Value : 0.03256

            Sensitivity : 0.8061
            Specificity : 0.6492
         Pos Pred Value : 0.7516
         Neg Pred Value : 0.7177
             Prevalence : 0.5684
         Detection Rate : 0.4582
   Detection Prevalence : 0.6096
      Balanced Accuracy : 0.7277

       'Positive' Class : 1
```

**Image Source** – Computed R Output

# Comparison of models

| | Model 1 (Logit) | Model 2 (Logit) | Model 3 (Logit) | Model 1 (Probit) | Model 2 (Probit) | Model 3 (Probit) |
|---|---|---|---|---|---|---|
| Intercept | -0.255 (0.21) | -2.885*** (0.49) | 1.1532 (0.7420) | -0.1283 (0.1299) | -1.7175*** (0.28) | 0.7645. (0.4394) |
| nwifeinc | -0.012. (0.007) | -0.02*** (0.00) | -0.0199* (0.0082) | -0.007. (0.004) | -0.0172*** (0.004) | -0.0113* (0.0048) |
| educ | | 0.25*** (0.041) | 0.2233*** (0.0429) | | 0.1512*** (0.024) | 0.1315*** (0.0250) |
| exper | 0.096*** (0.012) | 0.089*** (0.012) | 0.117*** (0.013) | 0.054*** (0.006) | 0.0504*** (0.087) | 0.0691*** (0.0075) |
| age | | | -0.09*** (0.013) | | | -0.0579*** (0.0077) |
| kidslt6 | -0.65*** (0.15) | -0.854*** (0.17) | -1.463*** (0.200) | -0.408*** (0.09) | -0.5221*** (0.1008) | -0.8862*** (0.1166) |

# Conclusion

- Women employment is influenced by both factors i.e., direct and indirect.
- Most influencing factor for a women to have employment is number kids less than 6 years(approx. 70-80%).
- Other more influencing factors are direct factors like her education(approx. 15-25%) and age(approx. 10-20%).
- Indirect factors like her family income(approx. 1-3%) and family education influence a women employment but less significantly.
- Model 3 are accepted due to better value of its accuracy.
- Based on the prediction of above models, the probability of women employed are 0.57

**Image Source** – Computed R Output

*Thank you*