STAT 429 Group Project:
Labour Force Participation

Andrew, Joseph, Jonathan

Dataset Description

Cross-section data from 1976 Panel Study of Income Dynamics (PSID) based on data for the previous year, 1975.

Labour force participation amongst married women:
0 - “no”
1 - “yes”

Research Question

Aside from obvious factors that have an effect on labour force participation in married women, such as years of experience. What other factors from our data set are significant?

Overview of Experiment:

- Initial Model (m1)
- Interpretive Model (m2)
- Predictive Model (m3)

Response and Predictor Variables

Response Variable:

Participation (Factor) = Did the individual participate in the labour force? (0 “no”, 1 “yes”)

Predictor Variables:

youngkids = Number of children less than 6 years old in household

oldkids = Number of children between 6 and 18 in household

age = wife’s age in years

College (Factor) = Did the individual (wife) attend college? (0 “no”, 1 “yes”)

Education = wife’s education in years

Experience = actual years of wife’s previous labour market experience.

Tax = marginal tax rate facing the wife

hcollege (Factor) = did the individual’s husband attend college? (0 “no”, 1 “yes”)

hhours = husband’s hours worked in 1975

hage = husband’s age in years

heducation = husband’s education in years

hwage = husband’s wage, in 1975 dollars

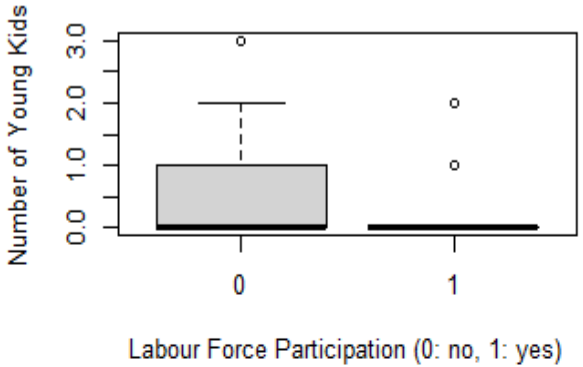
fincome = Family income, in 1975 dollars

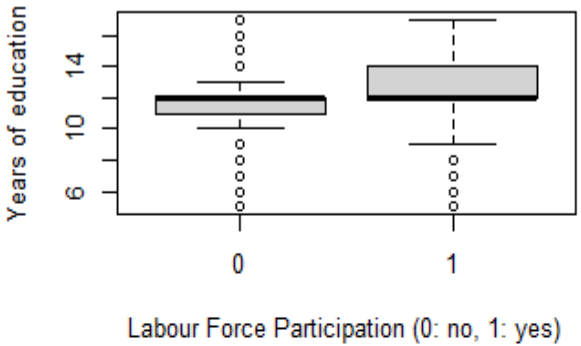
Meducation = wife’s mother’s educational attainment, in years

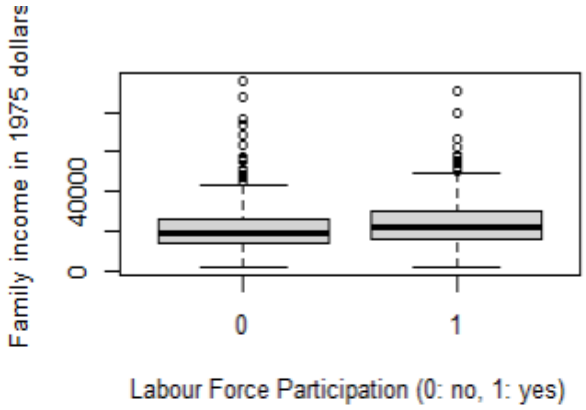
Feducation = wife’s father’s educational attainment, in years

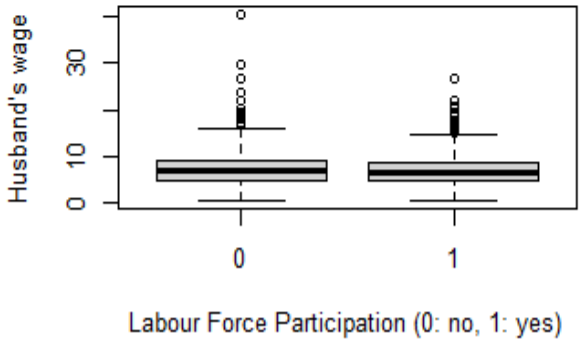
Unemp = unemployment rate in county of residence, in percentage points

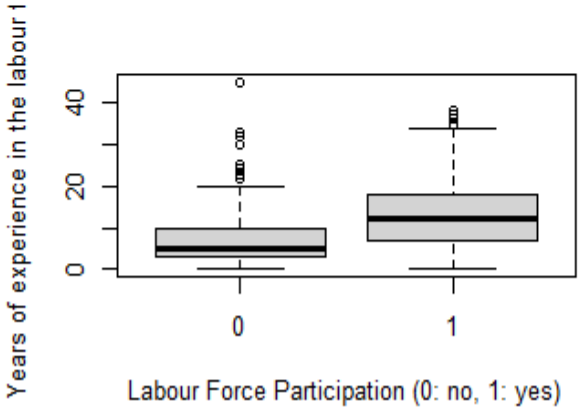
City (Factor) = Does the individual live in a large city? (0 “no”, 1 “yes”)


Number of Young Kids
0.0
1.0
2.0
3.0
0
1
Labour Force Participation (0: no, 1: yes)


Years of education
6
10
14
0
1
Labour Force Participation (0: no, 1: yes)


Family income in 1975 dollars
0
40000
0
1
Labour Force Participation (0: no, 1: yes)


Husband's wage
0
10
30
0
1
Labour Force Participation (0: no, 1: yes)


Years of experience in the labour f
0
20
40
0
1
Labour Force Participation (0: no, 1: yes)

Initial Model (m1)

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.643e+01  2.930e+00   5.609 2.03e-08 ***
youngkids   -1.330e+00  2.118e-01  -6.282 3.33e-10 ***
oldkids      1.602e-01  8.094e-02   1.979 0.047811 *
age         -9.396e-02  2.603e-02  -3.610 0.000306 ***
education    1.864e-01  7.064e-02   2.638 0.008334 **
hhours      -1.279e-03  2.124e-04  -6.023 1.72e-09 ***
hage        -1.208e-02  2.488e-02  -0.486 0.627318
heducation  -5.370e-02  5.859e-02  -0.917 0.359358
hwage       -3.441e-01  4.995e-02  -6.888 5.64e-12 ***
fincome      3.224e-05  1.824e-05   1.768 0.077070 .
tax         -1.389e+01  3.009e+00  -4.617 3.89e-06 ***
meducation  -1.798e-03  3.461e-02  -0.052 0.958574
city         2.133e-02  2.067e-01   0.103 0.917815
feducation  -1.832e-03  3.329e-02  -0.055 0.956113
unemp       -1.005e-02  3.017e-02  -0.333 0.739169
experience   1.159e-01  1.444e-02   8.030 9.74e-16 ***
college      1.655e-01  3.274e-01   0.505 0.613245
hcollege    -6.738e-03  3.337e-01  -0.020 0.983891
---
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:  736.76  on 735  degrees of freedom
AIC: 772.76

Number of Fisher Scoring iterations: 5
```

```
Call:
glm(formula = participation ~ youngkids + oldkids + age + education +
    hhours + hage + heducation + hwage + fincome + tax + meducation +
    city + feducation + unemp + experience + college + hcollege,
    family = "binomial", data = df)
```

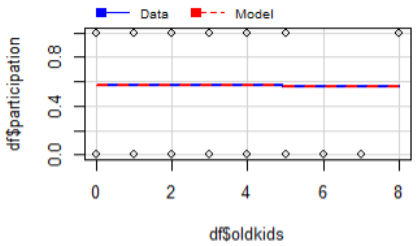
Interpretive Model (m2): Variable Selection

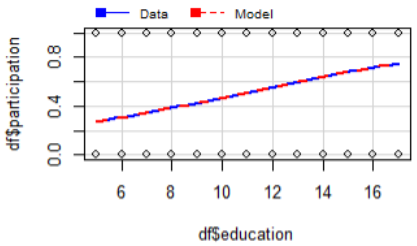
```
Step:  AIC=790.07
df$participation ~ youngkids + oldkids + age + education + hhours +
    hwage + tax + experience

             Df Deviance    AIC
<none>            742.52 790.07
- oldkids     1   748.51 790.78
- education   1   757.21 799.47
- hhours      1   785.44 827.70
- youngkids   1   786.85 829.12
- age         1   789.65 831.91
- hwage       1   807.23 849.50
- tax         1   807.52 849.78
- experience  1   819.61 861.87
```

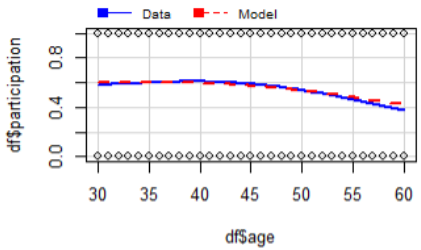
Interpretive Model (m2)

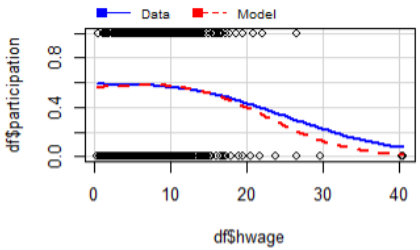
```
Call:
glm(formula = df$participation ~ oldkids + education + hhours +
    youngkids + age + hwage + tax + experience, family = "binomial",
    data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6565  -0.8166   0.3410   0.7664   2.8698

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.842e+01  2.475e+00   7.442 9.95e-14 ***
oldkids      1.901e-01  7.822e-02   2.430 0.015083 *
education    1.724e-01  4.608e-02   3.743 0.000182 ***
hhours      -1.274e-03  2.084e-04  -6.114 9.72e-10 ***
youngkids   -1.294e+00  2.065e-01  -6.266 3.71e-10 ***
age         -9.738e-02  1.500e-02  -6.490 8.58e-11 ***
hwage       -3.369e-01  4.720e-02  -7.138 9.50e-13 ***
tax         -1.730e+01  2.362e+00  -7.324 2.41e-13 ***
experience   1.141e-01  1.424e-02   8.009 1.16e-15 ***
---
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:  742.52  on 744  degrees of freedom
AIC: 760.52

Number of Fisher Scoring iterations: 5
```

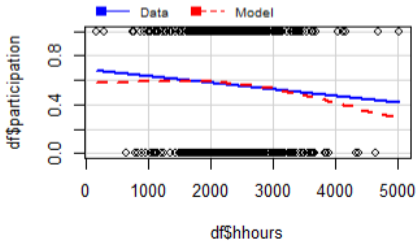
Interpretive Model (m2) : MMPS

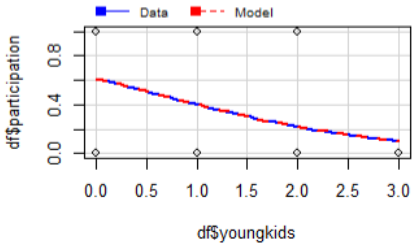


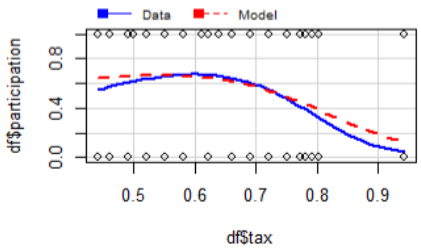


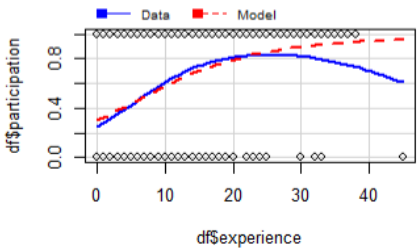




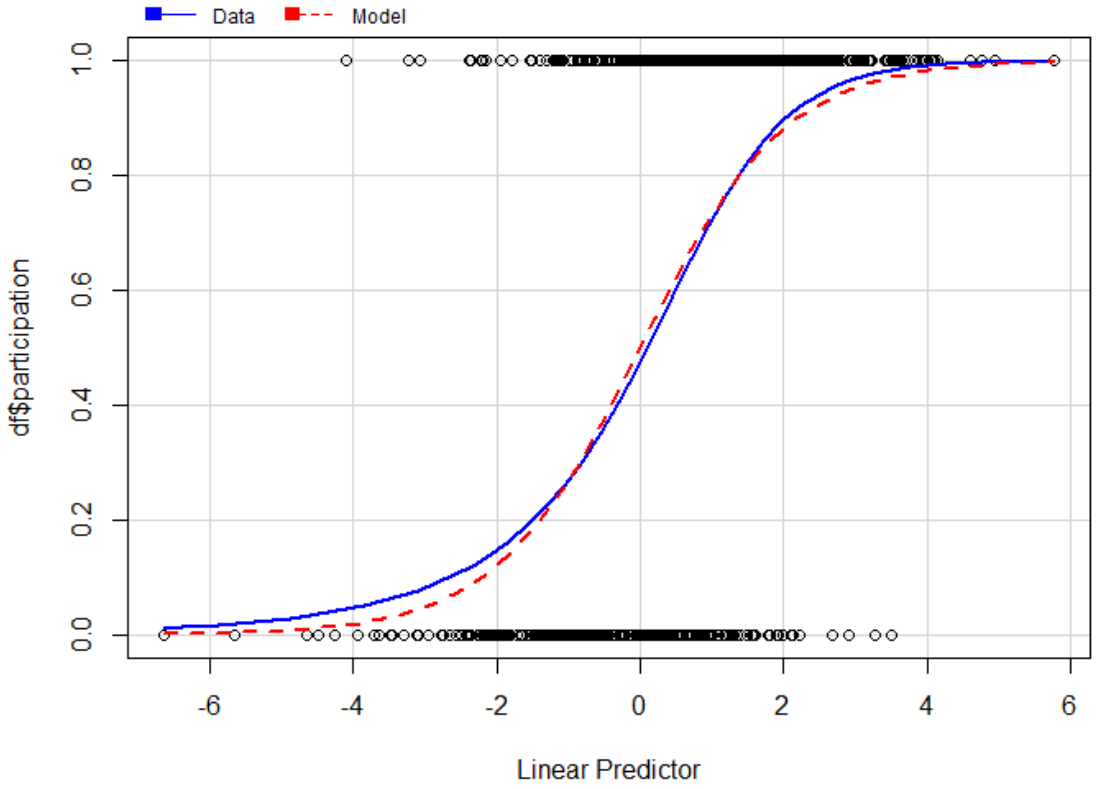








Interpretive Model (m2): Fit



Interpretive Model (m2): Anova

```
## Analysis of Deviance Table
##
## Model 1: df$participation ~ youngkids + oldkids + age + education + hhours +
##     hage + heducation + hwage + fincome + tax + meducation +
##     feducation + unemp + experience + city + college + hcollege
## Model 2: df$participation ~ oldkids + education + hhours + youngkids +
##     age + hwage + tax + experience
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       735     736.76
## 2       744     742.52 -9  -5.7678   0.7629
```

Interpretive Model (m2): Final Thoughts

```
vif(m2)
```

```
##    oldkids  education     hhours  youngkids        age      hwage        tax
##   1.348746   1.233898   1.770689   1.359423   1.734169   4.076285   4.068363
## experience
##   1.251825
```

Log odds of the predictors in the model for the average wife: **0.4047893**

Probability of the average wife being in the labour force: **0.5998378**

Per 1 year of additional experience, the odds in favour of the wife being in the labour force increased: **1.120833**

Predictive Model

The goal of this model is to minimize the AIC/BIC so that any prediction/confidence intervals are as tight as possible.

This will be done by:

- Attempting all possible transformation
- Selecting predictor variables with the all possible subset method

Predictive Model: Data Manipulation

Due to the nature of our data, certain variables have 0 as a possible value (not only the dummy ones).

Taking log transformations of this data will results in errors, therefore in order to be able to run the powerTransform() function we must replace the 0 with a small number

All zeros in; # of young kids, # of old kids, # of years of education husband, # of years of education wife, # of years of work experience wife. Were replaced with 10^-5

Predictive Model: Transformations

bcPower Transformations to Multinormality

	Est Power	Rounded Pwr	Wald Lwr Bnd	Wald Upr Bnd
Y1	-0.4236	-0.42	-0.4569	-0.3902
Y2	0.1570	0.16	0.1357	0.1784
Y3	0.3695	0.50	0.0628	0.6762
Y4	1.0902	1.00	0.8665	1.3138
Y5	0.6969	0.70	0.5674	0.8265
Y6	1.0095	1.00	0.7037	1.3154
Y7	1.3761	1.38	1.1562	1.5959
Y8	0.2861	0.33	0.2263	0.3459
Y9	0.2825	0.33	0.2191	0.3459
Y10	1.7286	1.73	1.4829	1.9743
Y11	0.7729	0.77	0.7039	0.8418
Y12	0.6924	0.69	0.6337	0.7511
Y13	0.6775	0.68	0.5132	0.8417
Y14	0.4321	0.43	0.3980	0.4662

The following transformations were recommended, and then all rounded powers were applied to the data

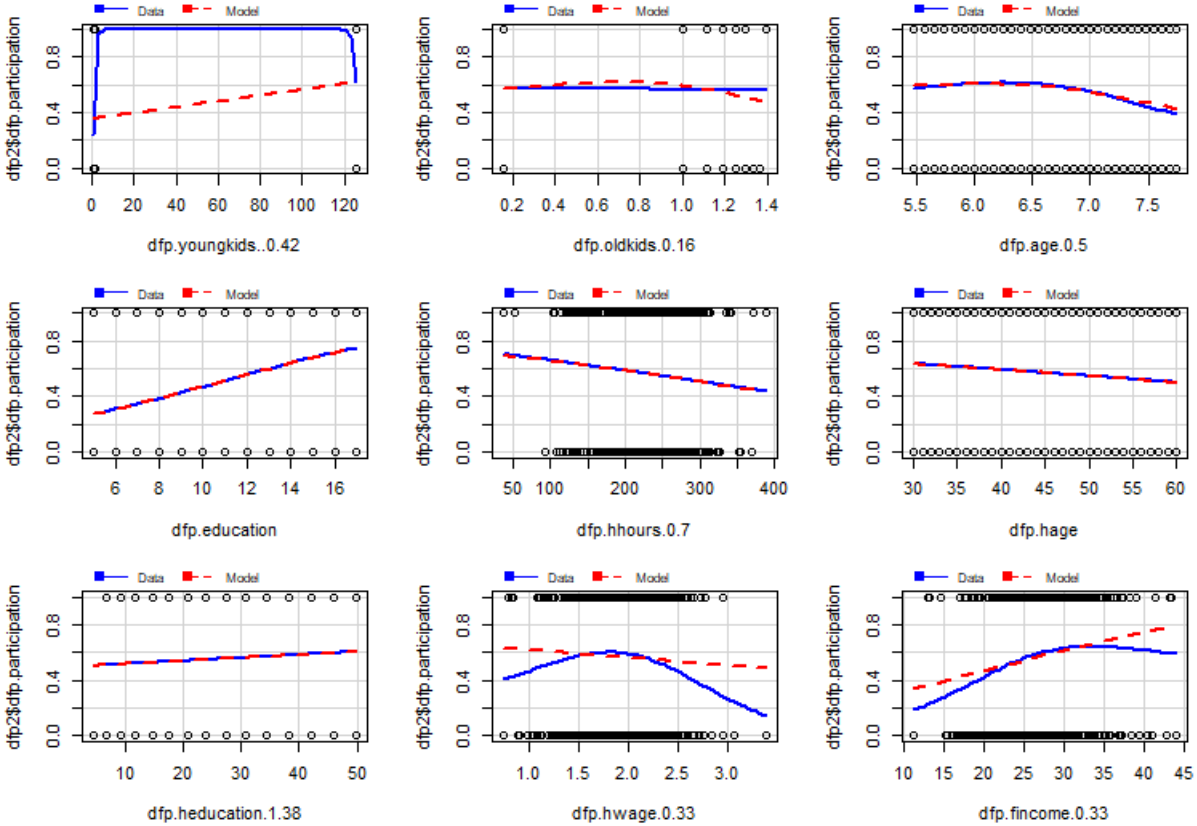
Note: the response variable and the dummy predictor variables were not transformed because of their binary nature.

Predictive Model: All Transformations

```
Call:
glm(formula = dfp2$dfp.participation ~ ., family = binomial,
    data = dfp2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4071  -0.8225   0.3404   0.7682   2.5013

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1029.75  on 752  degrees of freedom
Residual deviance:  741.05  on 735  degrees of freedom
AIC: 777.05

Number of Fisher Scoring iterations: 5
```

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             17.477191   3.253157   5.372 7.77e-08 ***
dfp.youngkids..0.42      0.012354   0.002149   5.749 9.00e-09 ***
dfp.oldkids.0.16         0.272936   0.241304   1.131 0.258018
dfp.age.0.5             -1.100252   0.333094  -3.303 0.000956 ***
dfp.education            0.143327   0.069760   2.055 0.039919 *
dfp.hhours.0.7          -0.020521   0.003241  -6.332 2.43e-10 ***
dfp.hage                -0.018885   0.024897  -0.759 0.448135
dfp.heducation.1.38     -0.020330   0.017226  -1.180 0.237928
dfp.hwage.0.33          -4.339057   0.609803  -7.116 1.12e-12 ***
dfp.fincome.0.33         0.139009   0.050631   2.746 0.006041 **
dfp.tax.1.73            -9.584409   2.275011  -4.213 2.52e-05 ***
dfp.meducation.0.77      0.006782   0.072238   0.094 0.925200
dfp.feducation.0.69      0.019761   0.090285   0.219 0.826750
dfp.unemp.0.68          -0.036060   0.086985  -0.415 0.678466
dfp.city                 0.034451   0.208677   0.165 0.868873
dfp.experience.0.43      0.895530   0.111457   8.035 9.38e-16 ***
dfp.college              0.119464   0.323211   0.370 0.711669
dfp.hcollege             0.067336   0.347708   0.194 0.846444
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Predictive Model: All Transformations MMP

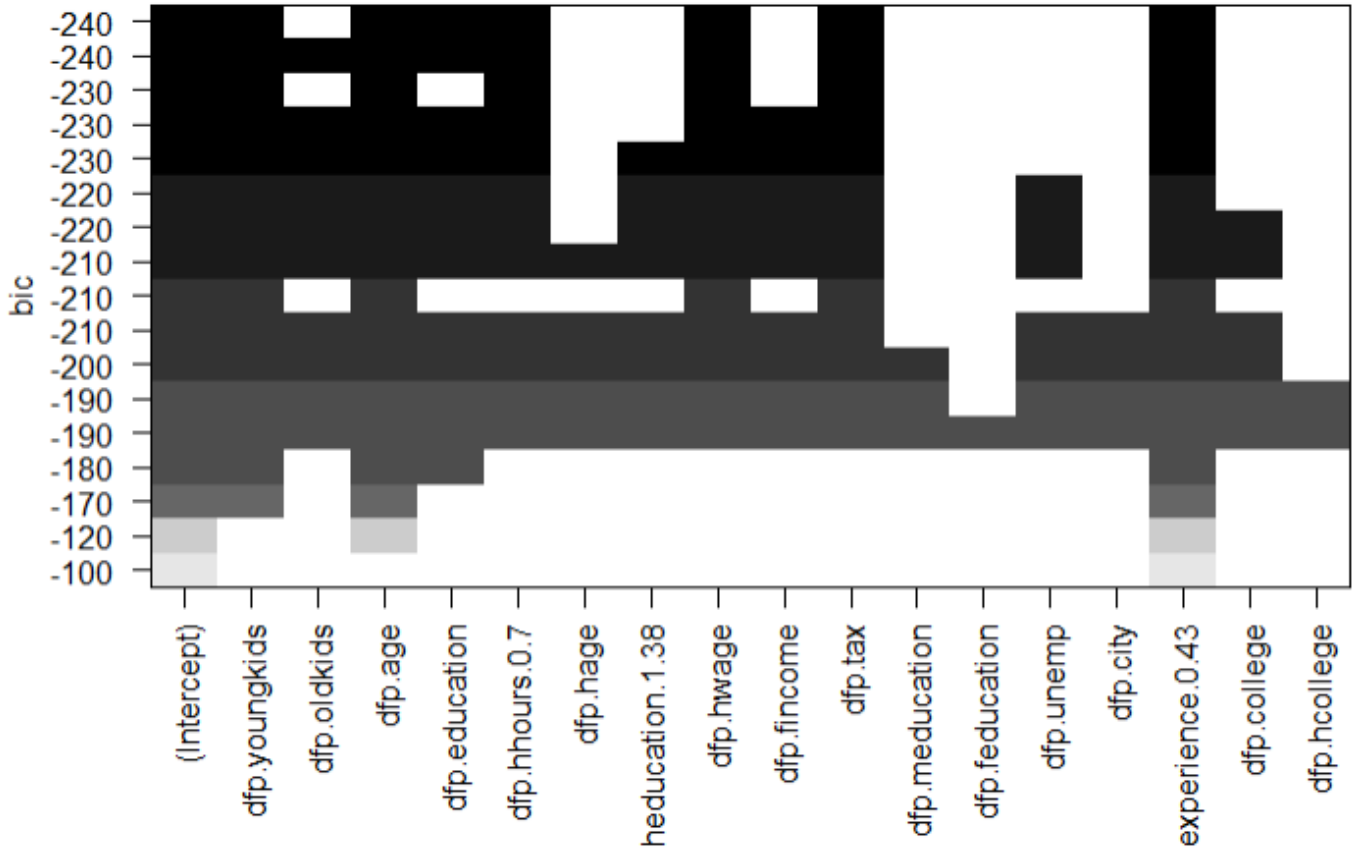


Predictive Model: Only good Transformations

```
Call:
glm(formula = dfp3$dfp.participation ~ ., family = binomial,
    data = dfp3)

Deviance Residuals:
    Min       1Q   Median       3Q      Max


(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1029.75  on 752  degrees of freedom
Residual deviance:  726.63  on 735  degrees of freedom
AIC: 762.63

Number of Fisher Scoring iterations: 5
```

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        1.716e+01  3.072e+00   5.584 2.35e-08 ***
dfp.youngkids     -1.303e+00  2.136e-01  -6.098 1.07e-09 ***
dfp.oldkids        1.616e-01  8.235e-02   1.962 0.049753 *
dfp.age           -8.645e-02  2.603e-02  -3.321 0.000898 ***
dfp.education      1.712e-01  7.068e-02   2.422 0.015421 *
dfp.hhours.0.7    -1.876e-02  3.107e-03  -6.037 1.57e-09 ***
dfp.hage          -1.343e-02  2.509e-02  -0.535 0.592370
dfp.heducation.1.38 -1.967e-02  1.738e-02  -1.132 0.257652
dfp.hwage         -3.554e-01  5.096e-02  -6.974 3.08e-12 ***
dfp.fincome        3.209e-05  1.832e-05   1.751 0.079947 .
dfp.tax           -1.465e+01  3.068e+00  -4.776 1.79e-06 ***
dfp.meducation    -4.379e-04  3.480e-02  -0.013 0.989959
dfp.feducation    -2.619e-03  3.370e-02  -0.078 0.938068
dfp.unemp         -2.213e-02  3.040e-02  -0.728 0.466691
dfp.city          -2.731e-02  2.090e-01  -0.131 0.896026
dfp.experience.0.43  9.293e-01  1.117e-01   8.318  < 2e-16 ***
dfp.college        2.006e-01  3.294e-01   0.609 0.542563
dfp.hcollege       5.887e-02  3.504e-01   0.168 0.866593
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

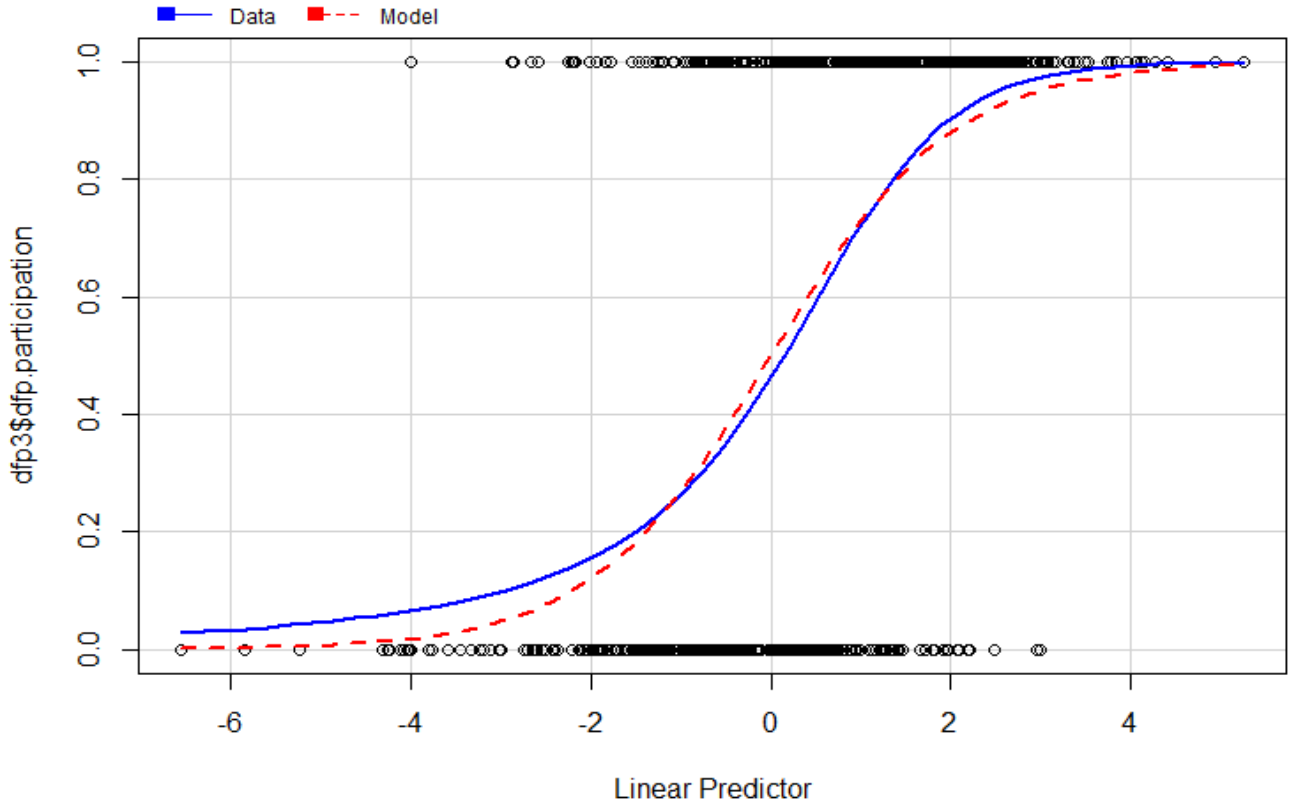
Predictive Model: Variable Selection



Predictive Model: Final

```
Call:
glm(formula = dfp3$dfp.participation ~ dfp3$dfp.youngkids + dfp3$dfp
    dfp3$dfp.education + dfp3$dfp.hhours.0.7 + dfp3$dfp.hwage +
    dfp3$dfp.tax + dfp3$dfp.experience.0.43, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4625  -0.8418   0.3567   0.7639   2.8362

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                18.960276   2.662338   7.122 1.07e-12 ***
dfp3$dfp.youngkids         -1.350759   0.204960  -6.590 4.39e-11 ***
dfp3$dfp.age               -0.104404   0.014137  -7.385 1.53e-13 ***
dfp3$dfp.education          0.150473   0.046129   3.262  0.00111 **
dfp3$dfp.hhours.0.7        -0.017461   0.002977  -5.865 4.50e-09 ***
dfp3$dfp.hwage             -0.333063   0.046776  -7.120 1.08e-12 ***
dfp3$dfp.tax              -16.927626   2.365442  -7.156 8.29e-13 ***
dfp3$dfp.experience.0.43    0.868865   0.107471   8.085 6.23e-16 ***
---
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:  739.3  on 745  degrees of freedom
AIC: 755.3

Number of Fisher Scoring iterations: 5
```

Predictive Model: Fit



Predictive Model: Final Thoughts

```
> vif(m4p)
      dfp3$dfp.youngkids              dfp3$dfp.age       dfp3$dfp.education      dfp3$dfp.hhours.0.7           dfp3$dfp.hwage
                1.356011                  1.569268                  1.241051                  1.743246                  3.963253
          dfp3$dfp.tax dfp3$dfp.experience.0.43
                3.945766                  1.189850
> |
```

Log odd of the average across all predictors: **0.3736192**

Probability of being in the labour force for average married woman: **0.5923332**

Percent of wives in the labour force: **0.5683931**

Model Comparisons: Confidence Intervals, AIC

Interpretive Model:

AIC: 760.52

```
> confint(m2)
Waiting for profiling to be done...
                    2.5 %        97.5 %
(Intercept)  13.688612532  2.340104e+01
oldkids       0.037673469  3.447680e-01
education     0.083371807  2.643177e-01
hhours       -0.001692782 -8.750819e-04
youngkids    -1.708781489 -8.978398e-01
age          -0.127423986 -6.853001e-02
hwage        -0.432399316 -2.470864e-01
tax         -22.054891040 -1.278542e+01
experience    0.086859718  1.427677e-01
```

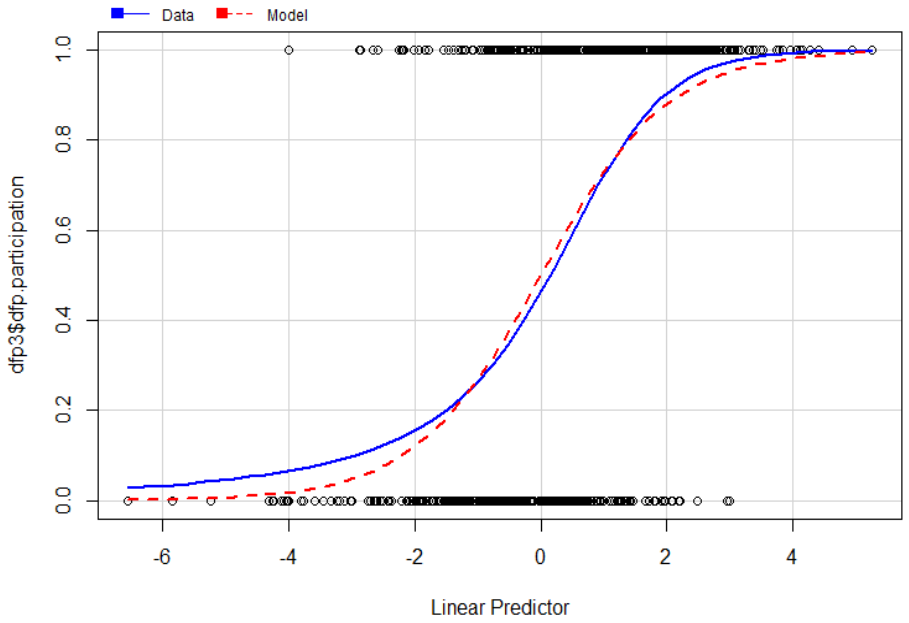
Predictive Model:

AIC: 755.3

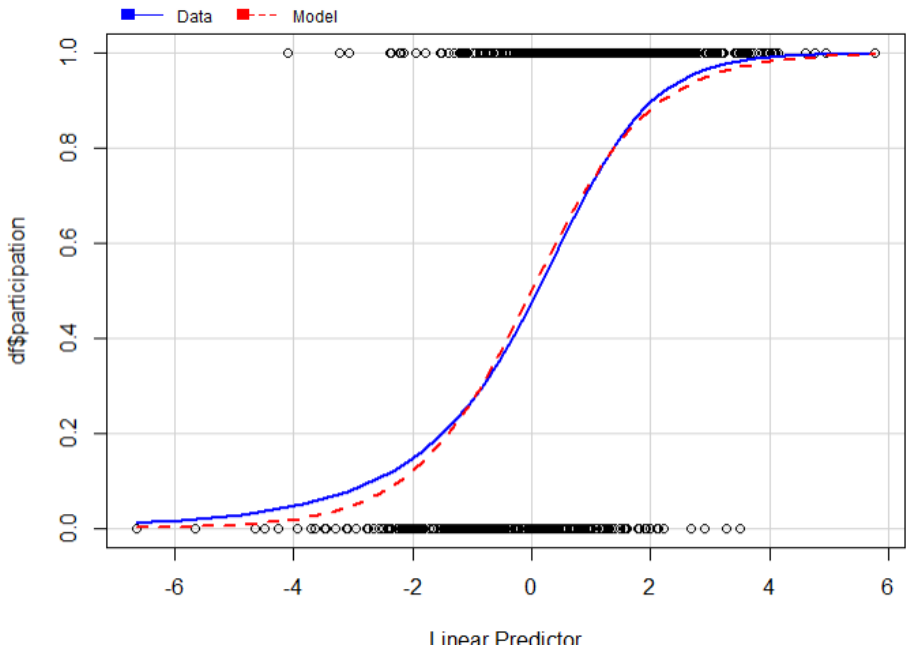
```
> confint(m4p)
Waiting for profiling to be done...
                                2.5 %       97.5 %
(Intercept)                13.87655969  24.32171186
dfp3$dfp.youngkids         -1.76285512  -0.95797988
dfp3$dfp.age               -0.13276239  -0.07727089
dfp3$dfp.education          0.06111768   0.24227343
dfp3$dfp.hhours.0.7        -0.02344960  -0.01176474
dfp3$dfp.hwage             -0.42765347  -0.24409761
dfp3$dfp.tax              -21.69884216 -12.41811990
dfp3$dfp.experience.0.43    0.66375017   1.08563069
```

Model Comparisons: Fit

Predictive Model (m3):



Interpretive Model (m2):



Conclusion

The extra transformation and reduced variable count of the **predictive model doesn't seem to be effective**.

The **interpretive model** has the upside of being able to understand it and has a **better** looking fit in the Marginal model plot..

$$participation = 18.42 + .1901(oldkids) + .1724(education) - .00124(hhours) - 1.294(youngkids) - 0.09738(age) - 0.3639(hwage) - 17.30(tax) + .01141(experience)$$

One thing to note: college is not present in both the interpretive model and the predictive model