DOM 207: IDSBRP
MINI PROJECT 4

INTRODUCTION

We, Aditya Arora (2010110038) and Sanjana Nair (2010110764) under Dr. Jaideep Ghosh as part of the DS Project Team for the Consulting firm CX, are presenting our findings in our project contracted out by the Government of India. The data provided contains 20 different parameters that affect women's participation in the labor force.We try to determine which factors are the primary determinants of women's participation in the labor force.

METHODOLOGY

Normalizing given data

We first try to determine whether the data of the different factors provided is distributed normally. For this we run shapiro-wilk's test in R. For the shapiro-wilk's test the null hypothesis is that the given distribution is normal. Hence, the alternative hypothesis becomes that the given distribution is not normal.

#HOURS

```
> shapiro.test(mwlab$hours)

        Shapiro-Wilk normality test

data:  mwlab$hours
W = 0.80675, p-value < 2.2e-16

> shapiro.test(log(mwlab$hours+1))

        Shapiro-Wilk normality test

data:  log(mwlab$hours + 1)
W = 0.73083, p-value < 2.2e-16
```

```
> shapiro.test(sqrt(mwlab$hours+1))

        Shapiro-Wilk normality test

data:  sqrt(mwlab$hours + 1)
W = 0.81961, p-value < 2.2e-16

> shapiro.test(1/(mwlab$hours+1))

        Shapiro-Wilk normality test

data:  1/(mwlab$hours + 1)
W = 0.63211, p-value < 2.2e-16
```

Not normally distributed

#AGE

```
> shapiro.test(mwlab$age)

        Shapiro-Wilk normality test

data:  mwlab$age
W = 0.96162, p-value = 3.829e-13

> shapiro.test(log(mwlab$age+1))

        Shapiro-Wilk normality test

data:  log(mwlab$age + 1)
W = 0.95948, p-value = 1.442e-13
```

```
> shapiro.test(sqrt(mwlab$age+1))

        Shapiro-Wilk normality test

data:  sqrt(mwlab$age + 1)
W = 0.96205, p-value = 4.686e-13

> shapiro.test(1/(mwlab$age+1))

        Shapiro-Wilk normality test

data:  1/(mwlab$age + 1)
W = 0.94597, p-value = 6.02e-16
```

Not normally distributed

#KIDSLT6

```
> shapiro.test(mwlab$kidslt6)

        Shapiro-Wilk normality test

data:  mwlab$kidslt6
W = 0.50285, p-value < 2.2e-16

> shapiro.test(log(mwlab$kidslt6+1))

        Shapiro-Wilk normality test

data:  log(mwlab$kidslt6 + 1)
W = 0.51449, p-value < 2.2e-16
```

```
> shapiro.test(sqrt(mwlab$kidslt6+1))

        Shapiro-Wilk normality test

data:  sqrt(mwlab$kidslt6 + 1)
W = 0.51197, p-value < 2.2e-16

> shapiro.test(1/(mwlab$kidslt6+1))

        Shapiro-Wilk normality test

data:  1/(mwlab$kidslt6 + 1)
W = 0.50928, p-value < 2.2e-16
```

#KIDSGE6

```
> shapiro.test(mwlab$kidsge6)

        Shapiro-Wilk normality test

data:  mwlab$kidsge6
W = 0.8629, p-value < 2.2e-16

> shapiro.test(log(mwlab$kidsge6+1))

        Shapiro-Wilk normality test

data:  log(mwlab$kidsge6 + 1)
W = 0.86464, p-value < 2.2e-16
```

```
> shapiro.test(sqrt(mwlab$kidsge6+1))

        Shapiro-Wilk normality test

data:  sqrt(mwlab$kidsge6 + 1)
W = 0.87938, p-value < 2.2e-16

> shapiro.test(1/(mwlab$kidsge6+1))

        Shapiro-Wilk normality test

data:  1/(mwlab$kidsge6 + 1)
W = 0.78797, p-value < 2.2e-16
```

Not normally distributed

Like the above three variable we observe how none of the variables are normally distributed as we observe that p value is smaller than 0.05 for each variable and their transformations, namely: log, square root, and multiplicative inverse.
This means that we reject the null hypothesis which states that the given distribution is normal. Hence our data is not normally distributed.

Next we try to normalize the data in the excel file (sheet=2). We do this by computing the mean and standard deviation using the functions average() and stdev.s() which computes the standard deviation for a sample.After this we take the log of the data and check if the data now becomes normal.

```
> shapiro.test(mwlab$hours)

        Shapiro-Wilk normality test

data:  mwlab$hours
W = 0.80675, p-value < 2.2e-16

> shapiro.test(mwlab$kidslt6)

        Shapiro-Wilk normality test

data:  mwlab$kidslt6
W = 0.50285, p-value < 2.2e-16

> shapiro.test(mwlab$kidsge6)

        Shapiro-Wilk normality test

data:  mwlab$kidsge6
W = 0.8629, p-value < 2.2e-16
```

```
> shapiro.test(mwlab$age)

        Shapiro-Wilk normality test

data:  mwlab$age
W = 0.96162, p-value = 3.829e-13

> shapiro.test(mwlab$edu)

        Shapiro-Wilk normality test

data:  mwlab$edu
W = 0.88634, p-value < 2.2e-16

> shapiro.test(mwlab$husage)

        Shapiro-Wilk normality test

data:  mwlab$husage
W = 0.97011, p-value = 2.771e-11
```

```
> shapiro.test(mwlab$husedu)

        Shapiro-Wilk normality test

data:  mwlab$husedu
W = 0.93541, p-value < 2.2e-16

> shapiro.test(mwlab$huswage)

        Shapiro-Wilk normality test

data:  mwlab$huswage
W = 0.85803, p-value < 2.2e-16
```

```
> shapiro.test(mwlab$faminc)

        Shapiro-Wilk normality test

data:  mwlab$faminc
W = 0.87015, p-value < 2.2e-16

> shapiro.test(mwlab$mtr)

        Shapiro-Wilk normality test

data:  mwlab$mtr
W = 0.92004, p-value < 2.2e-16

> shapiro.test(mwlab$momedu)

        Shapiro-Wilk normality test

data:  mwlab$momedu
W = 0.90474, p-value < 2.2e-16
```

```
> shapiro.test(mwlab$dadedu)

        Shapiro-Wilk normality test

data:  mwlab$dadedu
W = 0.90307, p-value < 2.2e-16

> shapiro.test(mwlab$unem)

        Shapiro-Wilk normality test

data:  mwlab$unem
W = 0.92283, p-value < 2.2e-16

> shapiro.test(mwlab$exper)

        Shapiro-Wilk normality test

data:  mwlab$exper
W = 0.92959, p-value < 2.2e-16
```

```
> shapiro.test(mwlab$nwifeinc)

	Shapiro-Wilk normality test

data:  mwlab$nwifeinc
W = 0.83882, p-value < 2.2e-16

> shapiro.test(mwlab$repwage)

	Shapiro-Wilk normality test

data:  mwlab$repwage
W = 0.77067, p-value < 2.2e-16

> shapiro.test(mwlab$hushrs)

	Shapiro-Wilk normality test

data:  mwlab$hushrs
W = 0.95941, p-value = 1.393e-13
```

```
> shapiro.test(mwlab$wage)

	Shapiro-Wilk normality test

data:  mwlab$wage
W = 0.71702, p-value < 2.2e-16
```

We observe that p value is smaller than 0.05 for each variable.
This means that we reject the null hypothesis which states that the given distribution is normal. Hence our data is not normally distributed.
Since we fail to normalize the data, we decide to proceed with the data at hand.

Removing Unnecessary Variables

Our first approach is to remove variables by removing the variable with the highest vif value at each step.

```
> model<- glm(inlf ~ hours + kidslt6 + kidsge6+ age + edu+wage+repwage+hushrs+husage+husedu+huswage+faminc+
medu+dadedu+unem+city+exper+nwifeinc,data=mwlab, family=binomial(link="logit"))
Warning messages:
1: glm.fit: algorithm did not converge
2: glm.fit: fitted probabilities numerically 0 or 1 occurred
> summary(model)

Call:
glm(formula = inlf ~ hours + kidslt6 + kidsge6 + age + edu +
    wage + repwage + hushrs + husage + husedu + huswage + faminc +
    mtr + momedu + dadedu + unem + city + exper + nwifeinc, family = binomial(link = "logit"),
    data = mwlab)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-9.705e-05  -7.160e-07   2.100e-08   2.100e-08   1.441e-04

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.837e+01  1.405e+05   0.000    1.000
hours        1.316e-01  8.505e+01   0.002    0.999
kidslt6     -2.101e+00  9.592e+03   0.000    1.000
kidsge6     -1.407e+00  6.581e+03   0.000    1.000
age         -1.822e-01  1.192e+03   0.000    1.000
edu          1.287e+00  3.678e+03   0.000    1.000
wage         3.707e+01  1.658e+04   0.002    0.998
repwage     -6.124e+00  7.493e+03  -0.001    0.999
hushrs       1.257e-03  9.946e+00   0.000    1.000
husage       9.281e-02  9.631e+02   0.000    1.000
husedu      -4.333e-01  2.996e+03   0.000    1.000
huswage      1.865e-01  2.386e+03   0.000    1.000
faminc      -2.359e-02  7.991e+00  -0.003    0.998
mtr          1.428e+01  1.085e+05   0.000    1.000
momedu      -2.953e-01  2.300e+03   0.000    1.000
dadedu       6.495e-01  2.181e+03   0.000    1.000
unem        -7.759e-02  1.665e+03   0.000    1.000
city        -5.635e-01  9.952e+03   0.000    1.000
exper       -2.058e-02  6.187e+02   0.000    1.000
nwifeinc     2.362e+01  8.044e+03   0.003    0.998

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1.0297e+03  on 752  degrees of freedom
Residual deviance: 9.1526e-08  on 733  degrees of freedom
AIC: 40

Number of Fisher Scoring iterations: 25

> bptest(model,data=mwlab)

        studentized Breusch-Pagan test

data:  model
BP = 258.83, df = 19, p-value < 2.2e-16

> vif(model)
      hours     kidslt6     kidsge6         age         edu        wage     repwage      hushrs      husage
  98.089227    3.109179    4.468067   16.382219    8.077956   30.708019   12.365955    5.818043   11.901958
     husedu     huswage      faminc         mtr      momedu      dadedu        unem        city       exper
   6.622104   17.109061 2255.340082   12.465296    8.229381    8.228021    3.290678    2.993287    3.912742
   nwifeinc
2268.339929
```

Hence we first remove faminc and nwifeinc.

```
> model<- glm(inlf ~ hours + kidslt6 + kidsge6+ age + edu+wage+repwage+hushrs+husage+husedu+huswage+mtr+momedu+da
dedu+unem+city+exper,data=mwlab, family=binomial(link="logit"))
Warning messages:
1: glm.fit: algorithm did not converge
2: glm.fit: fitted probabilities numerically 0 or 1 occurred
> summary(model)

Call:
glm(formula = inlf ~ hours + kidslt6 + kidsge6 + age + edu +
    wage + repwage + hushrs + husage + husedu + huswage + mtr +
    momedu + dadedu + unem + city + exper, family = binomial(link = "logit"),
    data = mwlab)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-1.294e-04  -1.326e-06   2.100e-08   2.100e-08   1.938e-04


Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.776e+01  7.775e+04   0.000    1.000
hours        1.091e-01  3.496e+01   0.003    0.998
kidslt6     -3.006e+00  6.863e+03   0.000    1.000
kidsge6     -1.375e+00  4.073e+03   0.000    1.000
age         -2.552e-01  7.428e+02   0.000    1.000
edu          9.925e-01  2.125e+03   0.000    1.000
wage         3.747e+01  1.003e+04   0.004    0.997
repwage     -7.453e+00  4.033e+03  -0.002    0.999
hushrs       1.532e-03  6.031e+00   0.000    1.000
husage       1.090e-01  5.840e+02   0.000    1.000
husedu      -3.804e-01  2.126e+03   0.000    1.000
huswage      1.958e-01  1.128e+03   0.000    1.000
mtr          8.724e+00  6.232e+04   0.000    1.000
momedu      -1.923e-01  1.551e+03   0.000    1.000
dadedu       6.016e-01  1.283e+03   0.000    1.000
unem        -1.141e-01  1.110e+03   0.000    1.000
city        -8.359e-01  6.607e+03   0.000    1.000
exper       -4.000e-02  4.975e+02   0.000    1.000


(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1.0297e+03  on 752  degrees of freedom
Residual deviance: 1.7668e-07  on 735  degrees of freedom
AIC: 36

Number of Fisher Scoring iterations: 25

> bptest(model,data=mwlab)

        studentized Breusch-Pagan test

data:  model
BP = 74.251, df = 17, p-value = 3.95e-09

> vif(model)
    hours   kidslt6   kidsge6       age       edu      wage   repwage    hushrs    husage    husedu   huswage
 4.258066  2.043798  3.112594 11.807778  6.045822  3.185209  2.669853  4.132421  8.578364  6.599457  7.012313
      mtr    momedu    dadedu      unem      city     exper
 6.220145  6.862823  5.302755  2.417444  2.486907  3.258683
```

Post this we notice that removing the two variables has generally reduced VIF values but not below 5 and none of the variables are statistically significant.

Now with trial and error we find the optimum model that gives us the most statistically significant variables.

```
> model<- glm(inlf ~kidslt6+kidsge6+age +edu+repwage+hushrs+husage+husedu+huswage+momedu+city+exper+unem+dadedu+m
tr,data=mwlab, family=binomial(link="logit"))
> summary(model)

Call:
glm(formula = inlf ~ kidslt6 + kidsge6 + age + edu + repwage +
    hushrs + husage + husedu + huswage + momedu + city + exper +
    unem + dadedu + mtr, family = binomial(link = "logit"), data = mwlab)
```

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.056e+01  3.010e+00   3.508 0.000452 ***
kidslt6     -1.126e+00  2.777e-01  -4.056 4.99e-05 ***
kidsge6      8.453e-02  9.967e-02   0.848 0.396399
age         -6.530e-02  3.245e-02  -2.012 0.044193 *
edu          2.064e-01  7.589e-02   2.720 0.006528 **
repwage      1.404e+00  1.396e-01  10.059  < 2e-16 ***
hushrs      -8.200e-04  2.557e-04  -3.207 0.001340 **
husage      -3.598e-03  3.179e-02  -0.113 0.909906
husedu      -2.202e-02  5.448e-02  -0.404 0.686071
huswage     -2.220e-01  5.624e-02  -3.947 7.90e-05 ***
momedu       3.509e-02  4.355e-02   0.806 0.420433
city        -2.168e-01  2.612e-01  -0.830 0.406448
exper        7.193e-02  1.692e-02   4.252 2.12e-05 ***
unem        -2.143e-02  3.778e-02  -0.567 0.570512
dadedu      -1.987e-02  4.197e-02  -0.473 0.635876
mtr         -1.121e+01  2.774e+00  -4.042 5.31e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```


```
(Dispersion parameter for binomial family taken to be 1)
```


```
    Null deviance: 1029.75  on 752  degrees of freedom
Residual deviance:  469.84  on 737  degrees of freedom
AIC: 501.84

Number of Fisher Scoring iterations: 7

> bptest(model,data=mwlab)

	studentized Breusch-Pagan test

data:  model
BP = 87.532, df = 15, p-value = 2.857e-12

> vif(model)
 kidslt6  kidsge6      age      edu  repwage   hushrs   husage   husedu  huswage   momedu     city    exper
1.391873 1.352369 5.349923 2.017918 1.049749 1.672375 4.900125 1.963621 3.678522 1.654752 1.178328 1.275045
    unem   dadedu      mtr
1.089195 1.586970 3.673711
```

Hence now we see that our vif values are all under 5 and we have a good number of statistically significant variables.

Hypothesis Testing

For Intercept:

Null Hypotheis: Coefficient for intercept=0

Alternative Hypothesis: Coefficient for intercept not equal to 0

We are able to reject the null hypothesis that intercept=0 as the p value <0.001

For kidslt6:

Null Hypotheis: Coefficient for kidslt6=0

Alternative Hypothesis: Coefficient for kidslt6 not equal to 0

We are able to reject the null hypothesis that kidslt6=0 as the p value is <0.001

For age:

Null Hypotheis: Coefficient for age=0

Alternative Hypothesis: Coefficient for age not equal to 0

We are able to reject the null hypothesis that age=0 as the p value < 0.05

For edu:

Null Hypotheis: Coefficient for edu=0

Alternative Hypothesis: Coefficient for edu not equal to 0

We are able to reject the null hypothesis that edu=0 as the p value is <0.001

For hushrs:

Null Hypotheis: Coefficient for hushrs=0

Alternative Hypothesis: Coefficient for hushrs not equal to 0

We are able to reject the null hypothesis that hsuhrs=0 as the p value <0.01

For huswage:

Null Hypotheis: Coefficient for huswage=0

Alternative Hypothesis: Coefficient for huswagenot equal to 0

We are able to reject the null hypothesis that huswage=0 as the p value <0.01

For repwage:

Null Hypotheis: Coefficient for repwage=0

Alternative Hypothesis: Coefficient for repwage not equal to 0

We are able to reject the null hypothesis that repwage=0 as the p value < 0.001

For exper:

Null Hypotheis: Coefficient for exper=0

Alternative Hypothesis: Coefficient for exper not equal to 0

We are able to reject the null hypothesis that exper=0 as the p value < 0.001

For mtr:

Null Hypotheis: Coefficient for mtr=0

Alternative Hypothesis: Coefficient for mtr not equal to 0

We are able to reject the null hypothesis that mtr=0 as the p value < 0.001

All other variable have p value>0.05 hence not significant.

OBSERVATION

Our dependent variable is Inlf and our control variable is City. Our intercept is extremely significant since p value is approximately 0. Our variables - Kidslt6, repwage, huswage,exper,mtr are also extremely significant. And hushrs,edu are highly significant and age is significant.The AIC value of our final model is far greater than the initial models we explored however it is only the final model which has any significant variables hence we feel it is a better choice to go with the final model instead of our initial model, since AIC is a comparative measure so this high AIC value might still be better than some even higher possible values while also having some statistically significant variables.

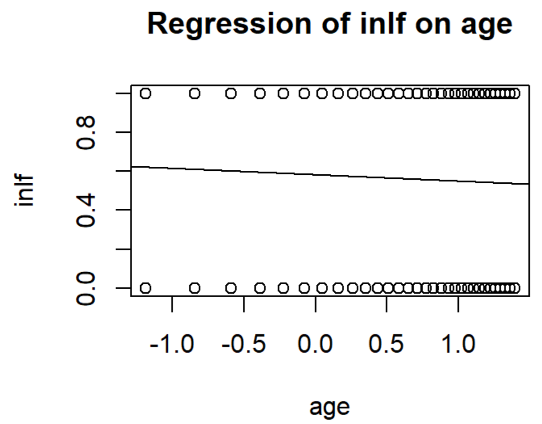
We have also run the breush-pagan test to check for heteroskedasticity.

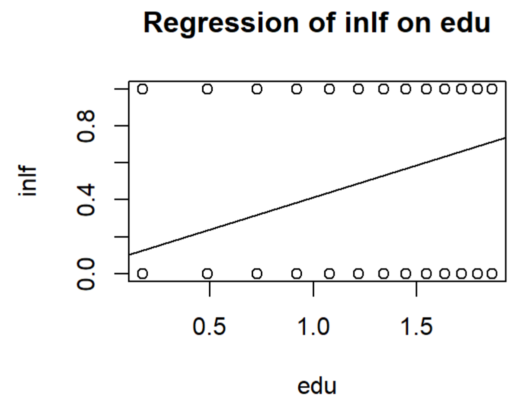
The null hypothesis for the test states that: Homoscedasticity is present

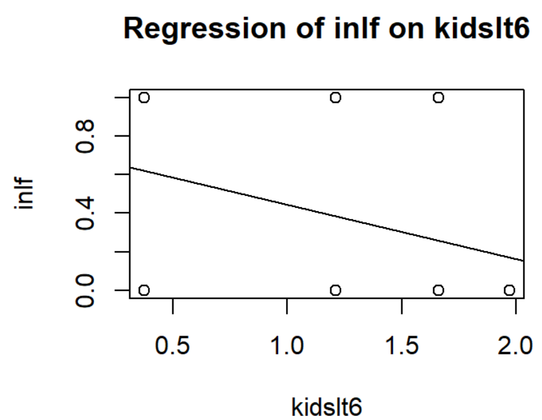
Hence the alternative hypothesis states that: heteroskedasticity is present.

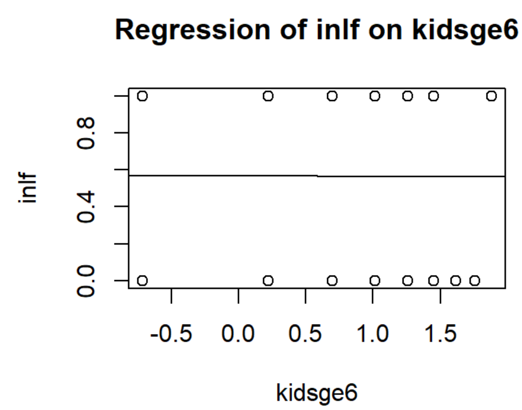
We get the p value for the breush-pagan test to be lower than 0.05 hence we reject the null hypothesis , hence heteroskedasticity is present in our model, however we are able to eliminate high vif values and bring all the vif values below 5, hence removing any significant multicollinearity. In general an acceptable logistic regression model can have some heteroskedasticity present but since the vif values are low enough we assume that the heteroskedasticity shouldn’t become a major issue.

Data Vizualization









We are unable to get proper graphs from which we can interpret the data, this is possible due to the fact that we were not able to normalize the data. Moreover since Heteroskedasticity is present, this might also interfere with the graphs.Hence the graph is not fitted for our regression model,but we reached a model that is statistically optimum.

Interpretation and Conclusion

```
> invlogit<-function(x)
+ {
+   1/(1+exp(-x))
+ }
> invlogit(model$coefficients)
 (Intercept)      kidslt6      kidsge6          age          edu      repwage       hushrs       husage 
0.9999740157 0.2448661263 0.5211200833 0.4836800894 0.5514201539 0.8028256554 0.4997950034 0.4991006031 
     husedu      huswage       momedu         city        exper         unem       dadedu          mtr 
0.4944949258 0.4447238051 0.5087708913 0.4460041248 0.5179753178 0.4946418631 0.4950316919 0.0000135441 
```

We take the invlogit of the coefficients to get the right coefficients which we can then interpret for our model. Repwage has the the highest coefficient value, since it represents the reported wage which will be the wage the women in labor force would expect.Mtr has the lowest coefficient as it represents the tax that the women who are employed have to pay on their salary but since they still receive salary it is still understandable that it would not affect much on their choice to be employed. Since employment is so necessary in today’s life it makes sense that every factor though however small would impact positively for someone to be chose being employed, and the fact that it is the labour industry or any other industry would not influence this significantly.