STAT536-HW1

Coco_Luo

2022-09-29

Question 1

Determine which variables should be excluded from the analysis up front. Examples include "hours", "wage", "lwage." These are variables that should not be seen as explanatory for the binary response "inlf".

- I choose to delete "hours", "wage", and "lwage" because those explanatory variables are directly related to the outcome variable. As long as those variables are greater than 0, we know 'inf' must be 1. Also, I noticed that "nwifeinc" is equal to (famimc - wage * hours)/1000, and the correlation between "nwifeinc" and "famimc" is as high as 0.94. Therefore, I would only ise one of them to build the model. I will remove "nwifeinc"

I have 17 explanatory variables remains, which are:

```
##  [1] "inlf"     "kidslt6"  "kidsge6"  "age"      "educ"     "repwage"
##  [7] "hushrs"   "husage"   "huseduc"  "huswage"  "faminc"   "mtr"
## [13] "motheduc" "fatheduc" "unem"     "city"     "exper"    "expersq"
```

Question 2

Determine if you need to take any transformation of the remaining explanatory variables. For example, "age" should be used as "log(age)", etc.

- We need to do transformations on variables that possess a non linear relationship with the response variable. And as I visualized each variabnle, I saw variables age and hushrs are factor variables which possess a non-linear relationship with the binary outcome, I decided to do a log transformation on those variables. Last but not the least, I standardized the explanatory variables so that they are more generalized.

```
labor2[,"age"] = log(labor2[,"age"])
labor2[,"husage"] = log(labor2[,"husage"])
labor2[,"hushrs"] = log(labor2[,"hushrs"])
```

```
data= labor2
for (i in 2:18){
  m = data[,i]
  data[,i] = (m-mean(m))/sd(m)
}
```

```
# correlations
for (i in 2:18){
  cat(colnames(data)[i], "=", cor(data[,1], data[,i]), "\n")
}
```

```
## kidslt6 = -0.2137493
## kidsge6 = -0.002424231
```

```
## age = -0.07226078
## educ = 0.1873528
## repwage = 0.6340485
## hushrs = -0.06160755
## husage = -0.06938293
## huseduc = 0.04591422
## huswage = -0.06947526
## faminc = 0.09889538
## mtr = -0.1448255
## motheduc = 0.09048973
## fatheduc = 0.05771841
## unem = -0.02873489
## city = -0.006167593
## exper = 0.3424847
## expersq = 0.2607407
```

Question 3

Determine an initial logistic regression M_0 using a heuristic argument of your choice. You could include the explanatory variables with the highest absolute correlation with the response. You can come up with another idea if you like.

- I would like to start from the model with explanatory variables having highest absolute correlation with the response. The threshold I chose here is 0.1. Thus, my initial model is M_0:

  $$logit(P(inlf = 1)) = \beta_0 + \beta_1 \cdot kidslt6 + \beta_2 \cdot educ + \beta_3 \cdot repwage + \beta_4 \cdot mtr + \beta_5 \cdot exper + \beta_6 \cdot expersq$$

Question 4

Fit the model M_0 and make sure there are no numerical errors when the MLEs are calculated. Give a formula for the fitted model and discuss its validity in term of standardized residuals, fitted values, etc. Produce relevant plots.

- From the summary statistics of the initial model I determined in question 3, it seems that mtr and expersq are not statistically significant, so I will delete them from my current model. Now, my fitted model is M_0:

  $$logit(P(inlf = 1)) = 1.74 - 0.37kidslt6 + 0.47educ + 3.57repwage + 0.43exper$$

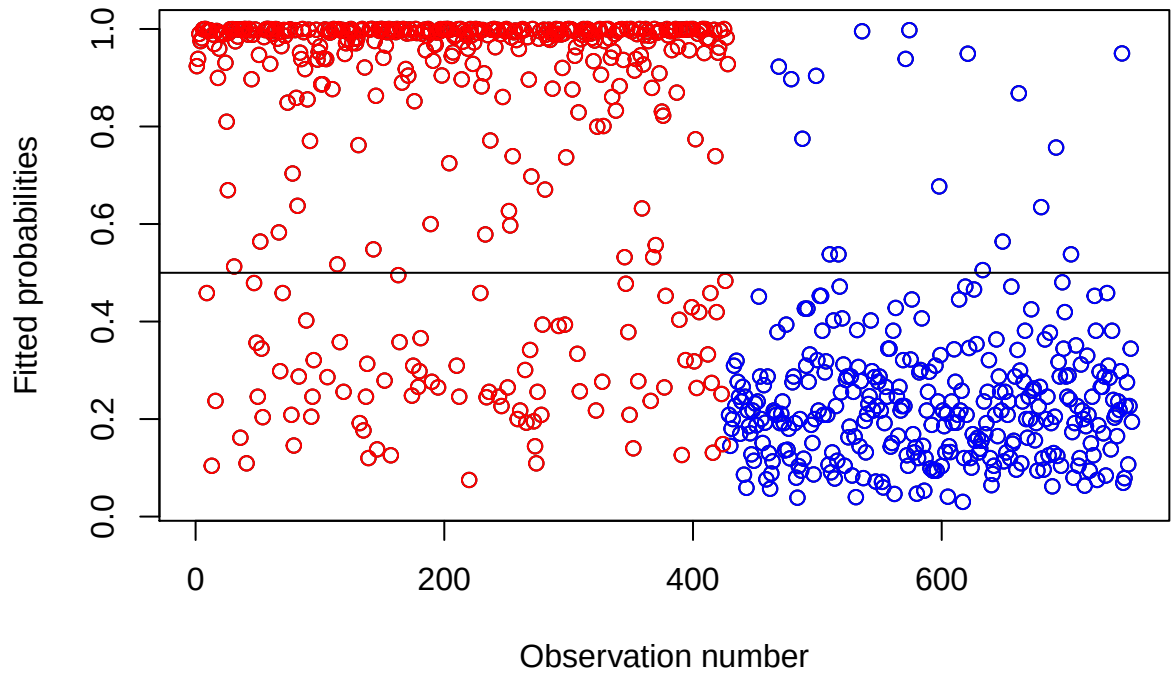
  Which makes sense as for each additional child younger than 6 years old a women has, the probability of her being in labor force decrease. And for more years of schooling she has, the probability of her being in labor force will increase.

```
inlf = data$inlf
# fit the model
M_0 <- glm(inlf ~ kidslt6+educ+repwage+exper,
         family = binomial(link = logit), data = data)
summary(M_0)
```

```
##
## Call:
## glm(formula = inlf ~ kidslt6 + educ + repwage + exper, family = binomial(link = logit),
##     data = data)
##
## Deviance Residuals:
```

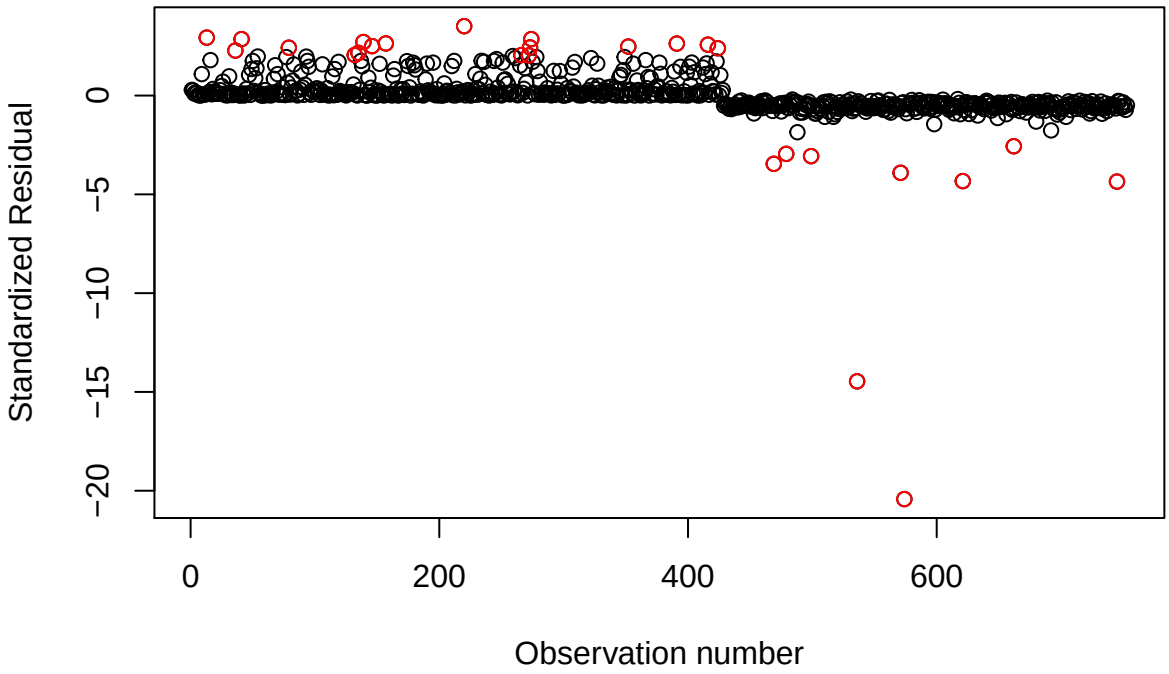
```
##     Min       1Q   Median       3Q      Max
## -3.4744  -0.6681   0.0195   0.2756   2.2757
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.7392     0.2358   7.375 1.65e-13 ***
## kidslt6      -0.3734     0.1220  -3.061 0.002206 **
## educ          0.4702     0.1201   3.914 9.07e-05 ***
## repwage       3.5650     0.3265  10.918  < 2e-16 ***
## exper         0.4279     0.1176   3.638 0.000275 ***
## ---
## 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:  512.18  on 748  degrees of freedom
## AIC: 522.18
##
## Number of Fisher Scoring iterations: 7
```

```r
# fitted values
idx = 1:length(data$inlf)
plot(idx,M_0$fitted.values,xlab="Observation number",ylab="Fitted probabilities")
points(idx[data$inlf==0],M_0$fitted.values[data$inlf==0],col="blue")
points(idx[data$inlf==1],M_0$fitted.values[data$inlf==1],col="red")
abline(h=0.5)
```




```r
#Standardized residuals
res = (data$inlf-M_0$fitted.values)/sqrt(M_0$fitted.values*(1-M_0$fitted.values))
#Plot it
plot(1:length(data$inlf),res,xlab="Observation number",ylab="Standardized Residual")
outlier = (1:length(data$inlf))[abs(res)>=2]
```

```
points(outlier,res[outlier],col='red')
```



For the first fitted values plot, values with outcome = 0 is marked as blue and values with outcome = 0 is marked as red. In addition, I plot out the outiers that is outside of the (-2,2) interval. There are about 27 outliers. Under the assumption of independence of the standardized residuals, I compute the probability of getting a value more extreme than the observed sum of the squares of the standardized residuals to be about 0, indicating our model can be further improved.

Question 5

Use the function `step` to improve M_0 in terms of AIC. Let M be the final model returned by `step`.

```
aic_M0 = M_0$deviance+2*length(coef(M_0))
bic_M0 = M_0$deviance+log(length(data$inlf))*length(coef(M_0))
aic_M0
```

```
## [1] 522.1823
```

```
bic_M0
```

```
## [1] 545.3026
```

```
library(leaps)
library(stats)
empty = glm(inlf~1,family=binomial(link=logit),data=data)
all = glm(inlf~.,family=binomial(link=logit),data=data)
M=step(empty, direction='forward', scope=formula(all), trace=0)
summary(M)
```

```
##
## Call:
## glm(formula = inlf ~ repwage + exper + educ + kidslt6 + age +
##     huswage + mtr + hushrs + faminc, family = binomial(link = logit),
##     data = data)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2046  -0.5565   0.0161   0.2439   2.5940
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.7930     0.2437   7.358 1.87e-13 ***
## repwage       3.3984     0.3360  10.115  < 2e-16 ***
## exper         0.5930     0.1346   4.404 1.06e-05 ***
## educ          0.4538     0.1367   3.319 0.000904 ***
## kidslt6      -0.6337     0.1436  -4.414 1.02e-05 ***
## age          -0.6439     0.1421  -4.532 5.83e-06 ***
## huswage      -1.0369     0.2331  -4.449 8.63e-06 ***
## mtr          -0.6412     0.2832  -2.264 0.023581 *
## hushrs       -0.5037     0.1491  -3.378 0.000731 ***
## faminc        0.3853     0.2604   1.480 0.138980
## ---
## 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:  470.65  on 743  degrees of freedom
## AIC: 490.65
##
## Number of Fisher Scoring iterations: 7
```

```
M$deviance+log(length(data$inlf))*length(coef(M))
```

```
## [1] 536.8938
```

The AIC of M_0 is 522 and BIC is 545. Then I use the step function to perform a forward and backward selection using the full model and the null model as the bounds.

The final model M:

$$logit(P(inlf = 1)) = 1.79 + 3.40 \cdot repwage + 0.60 \cdot exper + 0.45 \cdot educ - 0.64 \cdot kidslt6$$
$$-0.64 \cdot log(age) - 1.03 \cdot log(huswage) - 0.64 \cdot mtr - 0.50 \cdot log(hushrs) + 0.39 \cdot faminc$$

And we have an improved model M with AIC 490.91, and BIC 537.15.

Question 6

Consider all the models that are obtained by deleting one variable from M. Perform likelihood ratio tests to see whether any variables should be removed from M.

To see if each variable contained in the current model is related to the outcomne, I remove faminc and fit M_1 because from the last questions, only faminc is not statistically significant at level 0.01. A likelihood ratio test is used to see whether we can removed this variable from M or not.

```
M1 = update(M,. ~ . - faminc)
#drop1(M1,test="Chisq")
#1-pchisq(M1$deviance-M$deviance,1)
summary(M1)
```

```
##
## Call:
## glm(formula = inlf ~ repwage + exper + educ + kidslt6 + age +
##     huswage + mtr + hushrs, family = binomial(link = logit),
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2308  -0.5629   0.0173   0.2341   2.6882
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.7847     0.2435   7.330 2.31e-13 ***
## repwage       3.4027     0.3357  10.136  < 2e-16 ***
## exper         0.5649     0.1327   4.256 2.08e-05 ***
## educ          0.4478     0.1369   3.270  0.00108 **
## kidslt6      -0.6268     0.1428  -4.388 1.14e-05 ***
## age          -0.6234     0.1407  -4.432 9.33e-06 ***
## huswage      -0.9566     0.2280  -4.195 2.73e-05 ***
## mtr          -0.9079     0.2262  -4.015 5.95e-05 ***
## hushrs       -0.4659     0.1474  -3.161  0.00157 **
## ---
## 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:  472.84  on 744  degrees of freedom
## AIC: 490.84
##
## Number of Fisher Scoring iterations: 7
```

```
#M1$deviance+log(length(data$inlf))*length(coef(M1))
```

We see that the p-value associated with the removal of "faminc" is greater than 0.01 level, so we fail to reject the hypothesis that the coefficient of "faminc" in M is not zero. So I would consider removing "faminc" and obtain the model M1.All the variables in M1 are statistically significant. The AIC of M1 is 491.12, slightly larger than the AIC of M. But the BIC of M1 is 532.74 which is smaller than the BIC of M. Thus, I pick the smaller model with regard to BIC. M1 is shown as follows:

$$\begin{aligned} logit(P(inlf = 1)) = 1.79 + 3.41 \cdot repwage + 0.56 \cdot exper + 0.45 \cdot educ - 0.63 \cdot kidslt6 \\ -0.62 \cdot log(age) - 0.95 \cdot log(huswage) - 0.90 \cdot mtr - 0.47 \cdot hushrs \end{aligned}$$

I repeated this process for all the other variables, the p value shows that the removal is less than the threshold, and I fail to reject.

Question 7

Compute the BIC scores of the same models. Which model would you pick with respect to BIC?

```
# bic of same models - further delete variables
M2 = update(M1,. ~ . - repwage)
M2$deviance+M2$rank*log(length(inlf))
```

```
## [1] 803.9718
```

```
M2 = update(M1,. ~ . - exper)
M2$deviance+ M2$rank*log(length(inlf))
```

```
## [1] 544.6523
```

```
M2 = update(M1,. ~ . - educ)
M2$deviance+M2$rank*log(length(inlf))
```

```
## [1] 537.0734
```

```
M2 = update(M1,. ~ . - kidslt6)
M2$deviance+M2$rank*log(length(inlf))
```

```
## [1] 548.3787
```

```
M2 = update(M1,. ~ . - age)
M2$deviance+M2$rank*log(length(inlf))
```

```
## [1] 546.4634
```

```
M2 = update(M1,. ~ . - huswage)
M2$deviance+M2$rank*log(length(inlf))
```

```
## [1] 546.4884
```

```
M2 = update(M1,. ~ . - mtr)
M2$deviance+M2$rank*log(length(inlf))
```

```
## [1] 543.0502
```

```
M2 = update(M1,. ~ . - hushrs)
M2$deviance+M2$rank*log(length(inlf))
```

```
## [1] 536.2415
```

The results shows that no BIC is smaller than the BIC of M_1, I would still stick with M_1.

Question 8

Now calculate the Brier scores of the same models. Which model would you pick now? What is the error rate of your preferred model?

```
mean((inlf-M1$fitted.values)^2)
```

```
## [1] 0.0971737
```

```
 M3 = update(M1,. ~ . - repwage)
 mean((inlf-M3$fitted.values)^2)
```

```
## [1] 0.1638645
```

```
 M3 = update(M1,. ~ . - exper)
 mean((inlf-M3$fitted.values)^2)
```

```
## [1] 0.1026826
```

```
 M3 = update(M1,. ~ . - educ)
 mean((inlf-M3$fitted.values)^2)
```

```
## [1] 0.09972037
```

```
M3 = update(M1,. ~ . - kidslt6)
mean((inlf-M3$fitted.values)^2)
```

```
## [1] 0.1016333
```

```
M3 = update(M1,. ~ . - age)
mean((inlf-M3$fitted.values)^2)
```

```
## [1] 0.1020834
```

```
M3 = update(M1,. ~ . - huswage)
mean((inlf-M3$fitted.values)^2)
```

```
## [1] 0.1017878
```

```
M3 = update(M1,. ~ . - mtr)
mean((inlf-M3$fitted.values)^2)
```

```
## [1] 0.101455
```

```
M3 = update(M1,. ~ . - hushrs)
mean((inlf-M3$fitted.values)^2)
```

```
## [1] 0.09924785
```

The results shows that no Brier score is smaller than the Brier score of M_1, I would still stick with M_1.

Question 9

Draw conclusions about what you learned from your analysis. Give a formula for your final model and interpret the regression coefficients (an interpretation in term of log-odds is fine). Discuss the statistical significance of each variable selected in the model.

From the above analysis, I will choose M_1 as my final model where each variable selected in the model is statistically significant. And I refit it on the original data before centering and scaling for better interpretation.

$$logit(P(inlf = 1)) = 29.7 + 1.4 \cdot repwage + 0.07 \cdot exper + 0.2 \cdot educ$$
$$-1.2 \cdot kidslt6 - 3.24 \cdot log(age) - 0.2 \cdot huswage - 10.9 \cdot mtr - 1.6 \cdot hushrs$$

"repwage": for each unit of increase in "repwage", the odds of a woman being part of the labor force are expected to increase by a factor of $e^{(1.4)} = 4.1$, holding all other variables constant.

"exper": for each unit of increase in "exper", the odds of a woman being part of the labor force are expected to increase by a factor of $e^{(0.07)} = 1.1$, holding all other variables constant.

"educ": for each addition year of "education", the odds of a woman being part of the labor force are expected to increase by a factor of $e^{(0.2)} = 1.2$, holding all other variables constant.

"kidslt6": for each addition kids under 6, the odds of a woman being part of the labor force are expected to decrease by a factor of $e^{(1.2)} = 3.3$, holding all other variables constant.

"age": for each year of increase in "log(age)", the odds of a woman being part of the labor force are expected to decrease by a factor of $e^{(3.24)} = 25.5$, holding all other variables constant.

"huswage": for each unit of increase in "huswage", the odds of a woman being part of the labor force are expected to decrease by a factor of $e^{(0.2)} = 1.2$, holding all other variables constant.

"mtr": for each unit of increase in "mtr", the odds of a woman being part of the labor force are expected to decrease by a factor of $e^{(10.9)}$, holding all other variables constant.

"hushrs": for each hours of increase in "hushrs", the odds of a woman being part of the labor force are expected to decrease by a factor of $e^{(1.6)} = 5$, holding all other variables constant.