Logistic Regression

Roxanne Dimadi

2025-11-03

```r
#read the data; the file is saved in STATA format, so you need this library:
library(foreign)

# Load required libraries
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(janitor)
```

```
## Warning: package 'janitor' was built under R version 4.4.2
```

```
##
## Attaching package: 'janitor'
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```r
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 4.4.2
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```r
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```r
df <- read.dta('MROZ.dta');
```

```r
# Load your data
df %>%
  clean_names()  # makes column names snake_case
```

```r
# Quick overview

summary(df)
```

```
##       inlf             hours           kidslt6          kidsge6     
##  Min.   :0.0000   Min.   :   0.0   Min.   :0.0000   Min.   :0.000  
##  1st Qu.:0.0000   1st Qu.:   0.0   1st Qu.:0.0000   1st Qu.:0.000  
##  Median :1.0000   Median : 288.0   Median :0.0000   Median :1.000  
##  Mean   :0.5684   Mean   : 740.6   Mean   :0.2377   Mean   :1.353  
##  3rd Qu.:1.0000   3rd Qu.:1516.0   3rd Qu.:0.0000   3rd Qu.:2.000  
##  Max.   :1.0000   Max.   :4950.0   Max.   :3.0000   Max.   :8.000  
##                                                                    
##       age             educ            wage            repwage     
##  Min.   :30.00   Min.   : 5.00   Min.   : 0.1282   Min.   :0.00  
##  1st Qu.:36.00   1st Qu.:12.00   1st Qu.: 2.2626   1st Qu.:0.00  
##  Median :43.00   Median :12.00   Median : 3.4819   Median :0.00  
##  Mean   :42.54   Mean   :12.29   Mean   : 4.1777   Mean   :1.85  
##  3rd Qu.:49.00   3rd Qu.:13.00   3rd Qu.: 4.9708   3rd Qu.:3.58  
##  Max.   :60.00   Max.   :17.00   Max.   :25.0000   Max.   :9.98  
##                                  NA's   :325                      
##      hushrs         husage         huseduc         huswage       
##  Min.   : 175   Min.   :30.00   Min.   : 3.00   Min.   : 0.4121  
##  1st Qu.:1928   1st Qu.:38.00   1st Qu.:11.00   1st Qu.: 4.7883  
##  Median :2164   Median :46.00   Median :12.00   Median : 6.9758  
##  Mean   :2267   Mean   :45.12   Mean   :12.49   Mean   : 7.4822  
##  3rd Qu.:2553   3rd Qu.:52.00   3rd Qu.:15.00   3rd Qu.: 9.1667  
##  Max.   :5010   Max.   :60.00   Max.   :17.00   Max.   :40.5090  
##                                                                  
##      faminc           mtr            motheduc         fatheduc     
##  Min.   : 1500   Min.   :0.4415   Min.   : 0.000   Min.   : 0.000  
##  1st Qu.:15428   1st Qu.:0.6215   1st Qu.: 7.000   1st Qu.: 7.000  
##  Median :20880   Median :0.6915   Median :10.000   Median : 7.000  
##  Mean   :23081   Mean   :0.6789   Mean   : 9.251   Mean   : 8.809  
##  3rd Qu.:28200   3rd Qu.:0.7215   3rd Qu.:12.000   3rd Qu.:12.000  
##  Max.   :96000   Max.   :0.9415   Max.   :17.000   Max.   :17.000  
##                                                                    
##       unem             city            exper          nwifeinc      
##  Min.   : 3.000   Min.   :0.0000   Min.   : 0.00   Min.   :-0.02906  
##  1st Qu.: 7.500   1st Qu.:0.0000   1st Qu.: 4.00   1st Qu.:13.02504  
##  Median : 7.500   Median :1.0000   Median : 9.00   Median :17.70000  
##  Mean   : 8.624   Mean   :0.6428   Mean   :10.63   Mean   :20.12896  
##  3rd Qu.:11.000   3rd Qu.:1.0000   3rd Qu.:15.00   3rd Qu.:24.46600  
##  Max.   :14.000   Max.   :1.0000   Max.   :45.00   Max.   :96.00000  
##                                                                      
##      lwage            expersq    
##  Min.   :-2.0542   Min.   :   0  
##  1st Qu.: 0.8165   1st Qu.:  16  
```

```
##  Median : 1.2476   Median :  81
##  Mean   : 1.1902   Mean   : 178
##  3rd Qu.: 1.6036   3rd Qu.: 225
##  Max.   : 3.2189   Max.   :2025
##  NA's   :325
```

```r
# Check missing values
colSums(is.na(df))
```

```
##     inlf    hours  kidslt6  kidsge6      age     educ     wage  repwage
##        0        0        0        0        0        0      325        0
##   hushrs   husage  huseduc  huswage   faminc      mtr motheduc fatheduc
##        0        0        0        0        0        0        0        0
##     unem     city    exper nwifeinc    lwage  expersq
##        0        0        0        0      325        0
```

##Question 1: Determine which variables should be excluded from analysis

```r
#the wage of the non-working women is zero, hence the variables
#wage and lwage (columns 7 and 21) will contain missing values
#we eliminate these variables from the data

# colnames(df)

df <- df %>% select(-repwage)
df <- df %>% select(-lwage, -wage, -hours)
```

Question 2: Determine which variables need to be transformed

```r
#tranform some variables
df[,"age"] = log(df[,"age"])
df[,"hushrs"] = log(df[,"hushrs"])

#display the names of the variables
colnames(df)
```

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

Question 3: Determine intial logistic regression M0 inluding explanatory variables with highest absolute correlation.

```r
# Ensure inlf is numeric (0/1)
df$inlf <- as.numeric(df$inlf)

# Compute correlation of all numeric columns with inlf
correlations <- cor(df, use = "complete.obs")[, "inlf"]
```

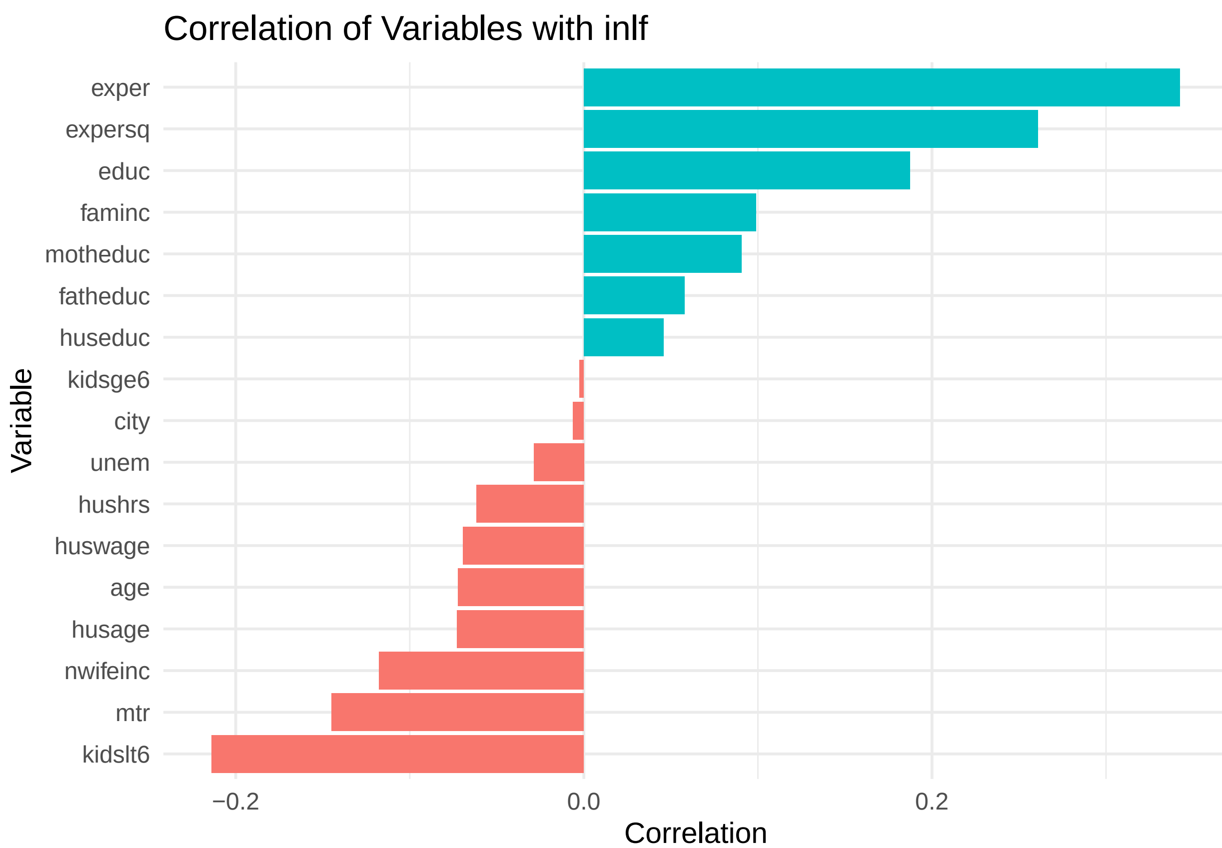
```r
# Sort by absolute correlation strength
correlations_sorted <- sort(abs(correlations), decreasing = TRUE)

# Display the results
correlations_sorted
```

```
##        inlf       exper     expersq     kidslt6        educ         mtr
## 1.000000000 0.342484655 0.260740738 0.213749303 0.187352846 0.144825469
##    nwifeinc      faminc    motheduc      husage         age     huswage
## 0.117597208 0.098895379 0.090489728 0.072820048 0.072260782 0.069475258
##      hushrs    fatheduc     huseduc        unem        city     kidsge6
## 0.061607550 0.057718407 0.045914222 0.028734889 0.006167593 0.002424231
```

```r
# make a pretty bar plot

data.frame(
  variable = names(correlations_sorted),
  correlation = correlations[ names(correlations_sorted) ]
) %>%
  filter(variable != "inlf") %>%
  ggplot(aes(x = reorder(variable, correlation), y = correlation, fill = correlation > 0)) +
  geom_col(show.legend = FALSE) +
  coord_flip() +
  labs(title = "Correlation of Variables with inlf",
       x = "Variable",
       y = "Correlation") +
  theme_minimal()
```



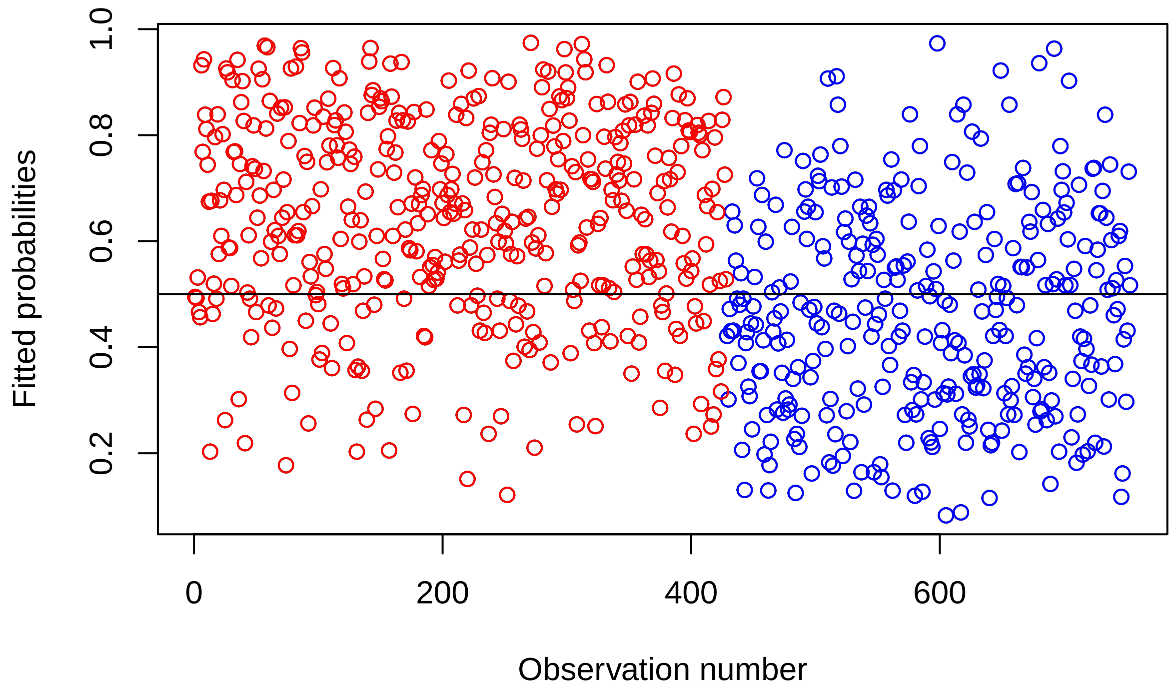
Question 4: Fit the Model M0 and make sure there are no numerical errors when MLEs are calculated. Give a formula for the fitted model and disucuss its validility in term of standardized residuals, fitted values, etc. Produce relevant plots

```
#fit a logistic regression with these variables
M0 = glm(inlf~kidslt6+educ+mtr+exper,family=binomial(link=logit),data=df)

formula(M0)
```

```
## inlf ~ kidslt6 + educ + mtr + exper
```

```
#make a plot of the fitted values
myind <- 1:nrow(df)
plot(myind,M0$fitted.values,xlab="Observation number",ylab="Fitted probabilities")
points(myind[df$inlf==0],M0$fitted.values[df$inlf==0],col="blue")
points(myind[df$inlf==1],M0$fitted.values[df$inlf==1],col="red")
abline(h=0.5)
```



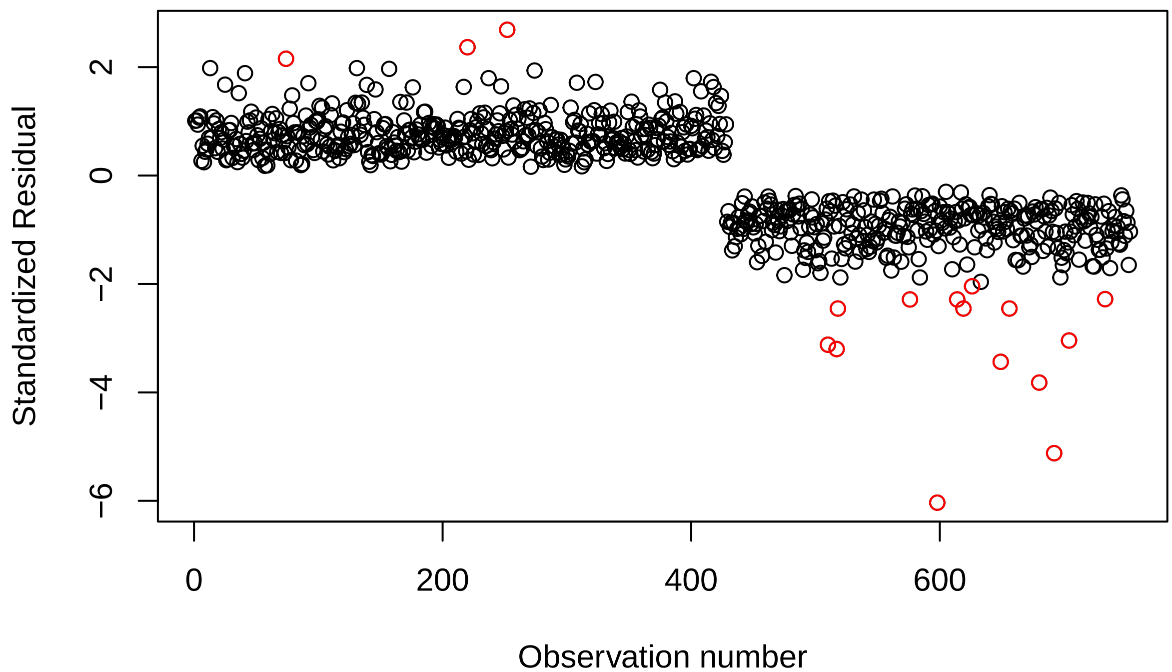

```r
summary(M0)
```

```
##
## Call:
## glm(formula = inlf ~ kidslt6 + educ + mtr + exper, family = binomial(link = logit),
##     data = df)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.50102    1.04784  -1.432    0.152
## kidslt6     -0.75643    0.16564  -4.567 4.95e-06 ***
## educ         0.17548    0.04143   4.236 2.27e-05 ***
## mtr         -1.68537    1.09275  -1.542    0.123
## exper        0.09624    0.01224   7.864 3.72e-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:  881.03  on 748  degrees of freedom
## AIC: 891.03
##
## Number of Fisher Scoring iterations: 4
```

```r
#determine the standardized residuals
myres = (df$inlf-M0$fitted.values)/sqrt(M0$fitted.values*(1-M0$fitted.values))


#make an index plot of standardized residuals against observation number
plot(1:length(df$inlf),myres,xlab="Observation number",ylab="Standardized Residual")

#the outliers are outside the (-2,2) interval
a = (1:length(df$inlf))[abs(myres)>=2]
points(a,myres[a],col='red')
```




```r
#calculate the p-value for the chisq test
1-pchisq(sum(myres^2),length(df$inlf)-length(coef(M0)))
```

```
## [1] 0.1103032
```

The fitted values plot shows the red points, which represent women in the labor force (infl = 1) and the blue points represent those not in the labor force (infl = 0). The y-axis shows the fitted probabilities of being in the labor force, with the horizontal line at 0.5 is the decision threshhold (prob = 0.5).

This model shows us that most red points are above 0.5, but the blue's are a bit mixed. If our model seperated the two groups reasonably well, we'd have clear red points above 0.5 and clear blue points under 0.5. Since our model doesn't do this, this shows that it misclassifies some cases.

The standardized residuals measure the difference between observed (0/1) and predicted probabilties. They should look roughly symmetric around 0.

My chi-squared goodness-of-fit test compares my model's residual deviance to a chi-square distribution.

The p-value $= 0.11 > 0.05$, so we fail to reject the null hypothesis that the model fits the data adequately.That means there's no strong evidence of lack of fit.

Question 5:

Use the function step to improve M0 in terms of AIC. Let M be the final model returned by step

```r
M <- step(M0,trace=TRUE)
```

```
## Start:  AIC=891.03
## inlf ~ kidslt6 + educ + mtr + exper
##
##           Df Deviance    AIC
## <none>         881.03 891.03
## - mtr      1   883.43 891.43
## - educ     1   899.94 907.94
## - kidslt6  1   903.91 911.91
## - exper    1   956.34 964.34
```

```r
AIC(M0)
```

```
## [1] 891.0329
```

```r
AIC(M)
```

```
## [1] 891.0329
```

```r
formula(M)
```

```
## inlf ~ kidslt6 + educ + mtr + exper
```

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

```r
M_no_kidslt6 <- glm(inlf ~ educ + mtr + exper, family = binomial, data = df)
M_no_educ    <- glm(inlf ~ kidslt6 + mtr + exper, family = binomial, data = df)
M_no_mtr     <- glm(inlf ~ kidslt6 + educ + exper, family = binomial, data = df)
M_no_exper   <- glm(inlf ~ kidslt6 + educ + mtr, family = binomial, data = df)

anova(M_no_kidslt6, M, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: inlf ~ educ + mtr + exper
## Model 2: inlf ~ kidslt6 + educ + mtr + exper
```

```
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       749     903.91
## 2       748     881.03  1   22.874 1.73e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(M_no_educ,    M, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: inlf ~ kidslt6 + mtr + exper
## Model 2: inlf ~ kidslt6 + educ + mtr + exper
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       749     899.94
## 2       748     881.03  1   18.904 1.375e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(M_no_mtr,     M, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: inlf ~ kidslt6 + educ + exper
## Model 2: inlf ~ kidslt6 + educ + mtr + exper
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       749     883.43
## 2       748     881.03  1   2.3986   0.1214
```

```r
anova(M_no_exper,   M, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: inlf ~ kidslt6 + educ + mtr
## Model 2: inlf ~ kidslt6 + educ + mtr + exper
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       749     956.34
## 2       748     881.03  1   75.305 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

'mtr' is not significant. The increase in deviance is small (around = 2.4), and the p-value (0.12) is well above 0.05. so mtr does not meaningfully improve the model and can be dropped.

Question 7:

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

```r
BIC(M)
```

```
## [1] 914.1532
```

```r
BIC(M_no_kidslt6)
```

```
## [1] 930.4032
```

```r
BIC(M_no_educ)
```

```
## [1] 926.4328
```

```r
BIC(M_no_mtr)
```

```
## [1] 909.9277
```

```r
BIC(M_no_exper)
```

```
## [1] 982.8343
```

With respect to BIC, the model without mtr (inlf ~ kidslt6 + educ + exper) is the best-fitting model because it has the lowest BIC (909.93).

BIC penalizes model complexity more heavily than AIC, so removing mtr simplifies the model without losing predictive power.

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

```r
 #computetheBrierscore
# (brier1= sum((TD-mylogit1$fitted.values)^2))
brier_M = mean((df$inlf - M$fitted.values)^2)
brier_no_kidslt6 = mean((df$inlf - M_no_kidslt6$fitted.values)^2)
brier_no_educ = mean((df$inlf - M_no_educ$fitted.values)^2)
brier_no_mtr = mean((df$inlf - M_no_mtr$fitted.values)^2)
brier_no_exper = mean((df$inlf - M_no_exper$fitted.values)^2)

brier_M
```

```
## [1] 0.1991956
```

```r
brier_no_kidslt6
```

```
## [1] 0.2056429
```

```r
brier_no_educ
```

```
## [1] 0.2041714
```

```
brier_no_mtr
```

```
## [1] 0.1997614
```

```
brier_no_exper
```

```
## [1] 0.2224028
```

The Brier scores for the models were all around 0.20, indicating moderate predictive performance. The full model (inlf ~ kidslt6 + educ + mtr + exper) had the lowest Brier score (0.1992), suggesting the best overall predictive accuracy. However, the model excluding mtr (Brier = 0.1998) performed almost identically while being more parsimonious. Therefore, from a practical standpoint, I would select the simpler model (inlf ~ kidslt6 + educ + exper) as it balances interpretability and predictive power.

```
error.rate1 <- mean((M_no_mtr$fitted.values > 0.5 & df$inlf == 0) | (M_no_mtr$fitted.values < 0.5 & df$i
error.rate1
```

```
## [1] 0.2961487
```

The model has an error rate of 0.296, meaning that it misclassifies about 29.6% of the women in the dataset. This corresponds to an overall classification accuracy of roughly 70.4%. While the model performs reasonably well, the remaining error suggests that factors beyond education, experience, tax rate, and young children also influence labor force participation.

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

```
formula (M_no_mtr)
```

```
## inlf ~ kidslt6 + educ + exper
```

```
summary(M_no_mtr)
```

```
##
## Call:
## glm(formula = inlf ~ kidslt6 + educ + exper, family = binomial,
##     data = df)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.94561    0.48431  -6.082 1.19e-09 ***
## kidslt6     -0.78804    0.16528  -4.768 1.86e-06 ***
## educ         0.20072    0.03825   5.248 1.54e-07 ***
## exper        0.09588    0.01222   7.846 4.29e-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:  883.43  on 749  degrees of freedom
## AIC: 891.43
##
## Number of Fisher Scoring iterations: 4
```

General logisitic equation formula: log(pi / 1-pi)=B0+B1x1i+B2x2i+…+Bkxki

For my specific model: log(pi/1-pi) = -2.946 - 0.788(kidslt6) + 0.201(educ) + 0.096(exper)

Each beta is measuring how a one-unit change in that predictor affects the log-odds of being in the labor force, holding the others constant. For example, for each addition child under six, the log-odds of being in the labor force decrease by 0.788, holding all other variables constant. To convert to odds, it e^-0.788 = 0.455. This means that for every extra small child, the odds of being in the labor force are about 45% of what they were before (a ~55% drop in odds). In terms of probability, having one more child under six makes a woman significantly less likely to be in the labor force. Statistically, her odds of working fall by about half, and her probability of working typically decreases by 15–20 percentage points, depending on her education and experience.

For the intercept: -2.9456, it means when kidslt=0, educ=0, and exper = 0, the log-odds of being in the labor force are -2.9456. I talked about kidslt6 coeffient in the former paragraph, and for the educ coefficient, each additional year of education increases the log-odds of being in the labor force by 0.2007. For exper coefficient, each additional year of work experience increases the log-odds of being in the labor force by 0.0959.

To summarize, the log-odds of a woman being in the labor force increase with her education and work experience but decrease with the presence of young children.

For statistical significance, the p-value tests whether the coefficeint for each predictor is statiscally different from 0. The null hypothesis (H0) says the variable has no effect on the probability of being in the labor force (B = 0).The alternative hypothesis (H1) says the variable does have an effect (B =/ 0).

If the p-value is less than 0.05, we reject the null hypothesis and conclude that the predictor is statistically significant. Each variable had a p value much greater than 0.05, concluding that all three predictors contribute to explaining a woman's labor-force participation.