Live Session - Week 2: Discrete Response Models Lecture 2

Jeffrey Yau

1/15/2017

Agenda

1. Q&A (estimated time: 5 minutes)
2. An overview of this lecture (estimated time: 10 minutes)
3. An extended example (estimated time: 65 minutes)
4. Discussion of the GLM parameter estiamtion algorithm: Iterated Reweighted Least Square (estimated time: 10 minutes)

1. Questions?

2. An Overivew of this Lecture

This lecture begins the study of logistic regression models, the most important special case of the generalized linear models (GLMs).

Why does classical linear regression models is not appropriate, from both statistical sense and practical application sense, to model a respone variable that is a continuous numeric variable.

Topics covered in this lecture include

- An introduction to binary response models and linear probability model, covering the formulation of forme and its advantages limitations of the latter
- Binomial logistic regression model
- The logit transformation and the logistic curve
- Statistical assumption of binomial logistic regression model
- Maximum likelihood estimation of the parameters and an overview of a numerical procedure used in practice
- Variance-Covariance matrix of the estimators
- Hypothesis tests for the binomial logistic regression model parameters
- The notion of deviance and odds ratios in the context of logistic regression models
- Probability of success and the corresponding confidence intervals in the context of logistic regression models
- Common non-linear transformation used in the context of binary dependent variable
- Visual assessment of the logistic regression model
- R functions for *binomial distribution*
- the *link function* translates from the scale of mean response to the scale of linear predictor.

$$\eta(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

With $\mu(\mathbf{x}) = E(y|\mathbf{x})$ being the conditional mean of the response, we have in GLM

$$g(\mu(\mathbf{x})) = \eta(\mu(\mathbf{x}))$$

To estimate the parameters of a GLM model, MLE is used. Because there is generally no closed-form solution, numerical procedure is needed. In the case of GLM, the *iteratively weighted least squares* procedure is used.

Recall that the probability mass function of the Binomial random variable is

$$P(W_j = w_j) = \binom{n_j}{w_j} \pi_j^{w_j} (1 - \pi_j)^{n_j - w_j}$$

where $w_j = 0, 1, \ldots, n_j$ where $j = 1, 2$

3. An extended example

Insert the function to *tidy up* the code when they are printed out

```
library(knitr)
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)
```

In conducting solving data science problem, always begin with the understanding of the underlying question; our first step is **NOT** to analyze the data. Suppose the question is *"Whether or not women who attended college had a higher labor force participation rate?"* Note that this was not Mroz (1987)'s objective of his paper. For the sake of learning to use logistic regression in answering a specific question, we stick with this question in this example.

Understanding the sample: Remember that this sample comes from *1976 Panel Data of Income Dynamics (PSID)*. PSID is one of the most popular dataset used by economists.

First, load the car library in order to use the Mroz dataset and understand the structure dataset.

Typical questions you should always ask include

- What are the number of variables (or "features" as they are typically called in data science in general and machine learning in specific) and number of observations (or "examlpes" in data science)?
- Are there any missing values?
- Are these variables sufficient for you to answer you questions?
- Note: in practice, you will likely query your data from many of tables and join them.

```
library(car)
require(dplyr)
```

```
## Loading required package: dplyr
```

```
## Warning: package 'dplyr' was built under R version 3.2.5
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:car':
##
##     recode
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
str(Mroz)
```

```
## 'data.frame':    753 obs. of  8 variables:
##  $ lfp : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ k5  : int  1 0 1 0 1 0 0 0 0 0 ...
##  $ k618: int  0 2 3 3 2 0 2 0 2 2 ...
##  $ age : int  32 30 35 34 31 54 37 54 48 39 ...
##  $ wc  : Factor w/ 2 levels "no","yes": 1 1 1 1 2 1 2 1 1 1 ...
```

```
##  $ hc  : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ lwg : num  1.2102 0.3285 1.5141 0.0921 1.5243 ...
##  $ inc : num  10.9 19.5 12 6.8 20.1 ...
```

```
glimpse(Mroz)  # glimpse can be use for any data.frame or table in R
```

```
## Observations: 753
## Variables: 8
## $ lfp  <fctr> yes, yes, yes, yes, yes, yes, yes, yes, yes, yes, yes, y...
## $ k5   <int> 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, ...
## $ k618 <int> 0, 2, 3, 3, 2, 0, 2, 0, 2, 2, 1, 1, 2, 2, 1, 3, 2, 5, 0, ...
## $ age  <int> 32, 30, 35, 34, 31, 54, 37, 54, 48, 39, 33, 42, 30, 43, 4...
## $ wc   <fctr> no, no, no, no, yes, no, yes, no, no, no, no, no, no, no...
## $ hc   <fctr> no, no, no, no, no, no, no, no, no, no, no, yes, yes, no...
## $ lwg  <dbl> 1.2101647, 0.3285041, 1.5141279, 0.0921151, 1.5242802, 1....
## $ inc  <dbl> 10.910001, 19.500000, 12.039999, 6.800000, 20.100000, 9.8...
```

```
# View(Mroz)

head(Mroz, 5)
```

```
##   lfp k5 k618 age  wc hc       lwg   inc
## 1 yes  1    0  32  no no 1.2101647 10.91
## 2 yes  0    2  30  no no 0.3285041 19.50
## 3 yes  1    3  35  no no 1.5141279 12.04
## 4 yes  0    3  34  no no 0.0921151  6.80
## 5 yes  1    2  31 yes no 1.5242802 20.10
```

```
some(Mroz, 5)
```

```
##     lfp k5 k618 age  wc  hc       lwg    inc
## 6   yes  0    0  54  no  no 1.5564855  9.859
## 106 yes  0    3  33 yes yes 0.3285041 21.000
## 349 yes  0    2  57  no  no 3.1555810 24.399
## 649  no  0    0  49  no  no 1.2357149 15.500
## 708  no  0    1  42 yes yes 1.2140751 24.106
```

```
tail(Mroz, 5)
```

```
##     lfp k5 k618 age  wc  hc       lwg    inc
## 749  no  0    2  40 yes yes 1.0828638 28.200
## 750  no  2    3  31  no  no 1.1580402 10.000
## 751  no  0    0  43  no  no 0.8881401  9.952
## 752  no  0    0  60  no  no 1.2249736 24.984
## 753  no  0    3  39  no  no 0.8532125 28.363
```

```
library(Hmisc)
```

```
## Loading required package: grid
## Loading required package: lattice
## Loading required package: survival
## Loading required package: Formula
## Loading required package: ggplot2
## Warning: package 'ggplot2' was built under R version 3.2.5
## Warning: replacing previous import by 'ggplot2::unit' when loading 'Hmisc'
```

```
## Warning: replacing previous import by 'ggplot2::arrow' when loading 'Hmisc'
## Warning: replacing previous import by 'scales::alpha' when loading 'Hmisc'
##
## Attaching package: 'Hmisc'
## The following objects are masked from 'package:dplyr':
##
##     combine, src, summarize
## The following objects are masked from 'package:base':
##
##     format.pval, round.POSIXt, trunc.POSIXt, units
```

```
describe(Mroz)
```

```
## Mroz
##
##  8  Variables      753  Observations
## ---------------------------------------------------------------------------
## lfp
##       n missing  unique
##     753       0       2
##
## no (325, 43%), yes (428, 57%)
## ---------------------------------------------------------------------------
## k5
##       n missing  unique    Info    Mean
##     753       0       4    0.47  0.2377
##
## 0 (606, 80%), 1 (118, 16%), 2 (26, 3%), 3 (3, 0%)
## ---------------------------------------------------------------------------
## k618
##       n missing  unique    Info    Mean
##     753       0       9    0.93   1.353
##
##             0   1   2   3  4  5 6 7 8
## Frequency 258 185 162 103 30 12 1 1 1
## %          34  25  22  14  4  2 0 0 0
## ---------------------------------------------------------------------------
## age
##       n missing  unique    Info    Mean     .05     .10     .25     .50
##     753       0      31       1   42.54    30.6    32.0    36.0    43.0
##     .75     .90     .95
##    49.0    54.0    56.0
##
## lowest : 30 31 32 33 34, highest: 56 57 58 59 60
## ---------------------------------------------------------------------------
## wc
##       n missing  unique
##     753       0       2
##
## no (541, 72%), yes (212, 28%)
## ---------------------------------------------------------------------------
## hc
##       n missing  unique
```

```
##     753       0       2
##
## no (458, 61%), yes (295, 39%)
## ---------------------------------------------------------------------------
## lwg
##       n missing  unique    Info    Mean     .05     .10     .25     .50
##     753       0     676       1   1.097  0.2166  0.4984  0.8181  1.0684
##     .75     .90     .95
##  1.3997  1.7600  2.0753
##
## lowest : -2.054 -1.823 -1.766 -1.543 -1.030
## highest:  2.905  3.065  3.114  3.156  3.219
## ---------------------------------------------------------------------------
## inc
##       n missing  unique    Info    Mean     .05     .10     .25     .50
##     753       0     621       1   20.13   7.048   9.026  13.025  17.700
##     .75     .90     .95
##  24.466  32.697  40.920
##
## lowest : -0.029  1.200  1.500  2.134  2.200
## highest: 77.000 79.800 88.000 91.000 96.000
## ---------------------------------------------------------------------------
```

```
summary(Mroz)
```

```
##   lfp            k5              k618            age          wc
##  no :325   Min.   :0.0000   Min.   :0.000   Min.   :30.00   no :541
##  yes:428   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:36.00   yes:212
##            Median :0.0000   Median :1.000   Median :43.00
##            Mean   :0.2377   Mean   :1.353   Mean   :42.54
##            3rd Qu.:0.0000   3rd Qu.:2.000   3rd Qu.:49.00
##            Max.   :3.0000   Max.   :8.000   Max.   :60.00
##    hc           lwg               inc
##  no :458   Min.   :-2.0541   Min.   :-0.029
##  yes:295   1st Qu.: 0.8181   1st Qu.:13.025
##            Median : 1.0684   Median :17.700
##            Mean   : 1.0971   Mean   :20.129
##            3rd Qu.: 1.3997   3rd Qu.:24.466
##            Max.   : 3.2189   Max.   :96.000
```

Descriptive statistical analysis of the data

Summary of the descriptive statistical analysis:

1. No variable in the data set has missnig value. (This is very unlikely in practice, but this is a clean dataset used in many academic studies.)
2. The response (or dependent) variable of interest, female labor force participation denoted as *lfp*, is a binary variable taking the type "factor". The sample proporation of participation is 57% (or 428 people in the sample).
3. There are 7 potential explanatory variables included in this data:

- number of kids below the age of 5
- number of kids between 6 and 18
- wife's age (in years)

- wife's college attendance
- husband's college attendance
- log of wife's estimated wage rate
- family income excluding the wife's wage ($1000)

All of them are potential determinants of wife's labor force participation, although I am concern using the wage rate (until I can learn more about this variable) because only those who worked have a wage rate. Of course, we should not think of this list as exhaustive.

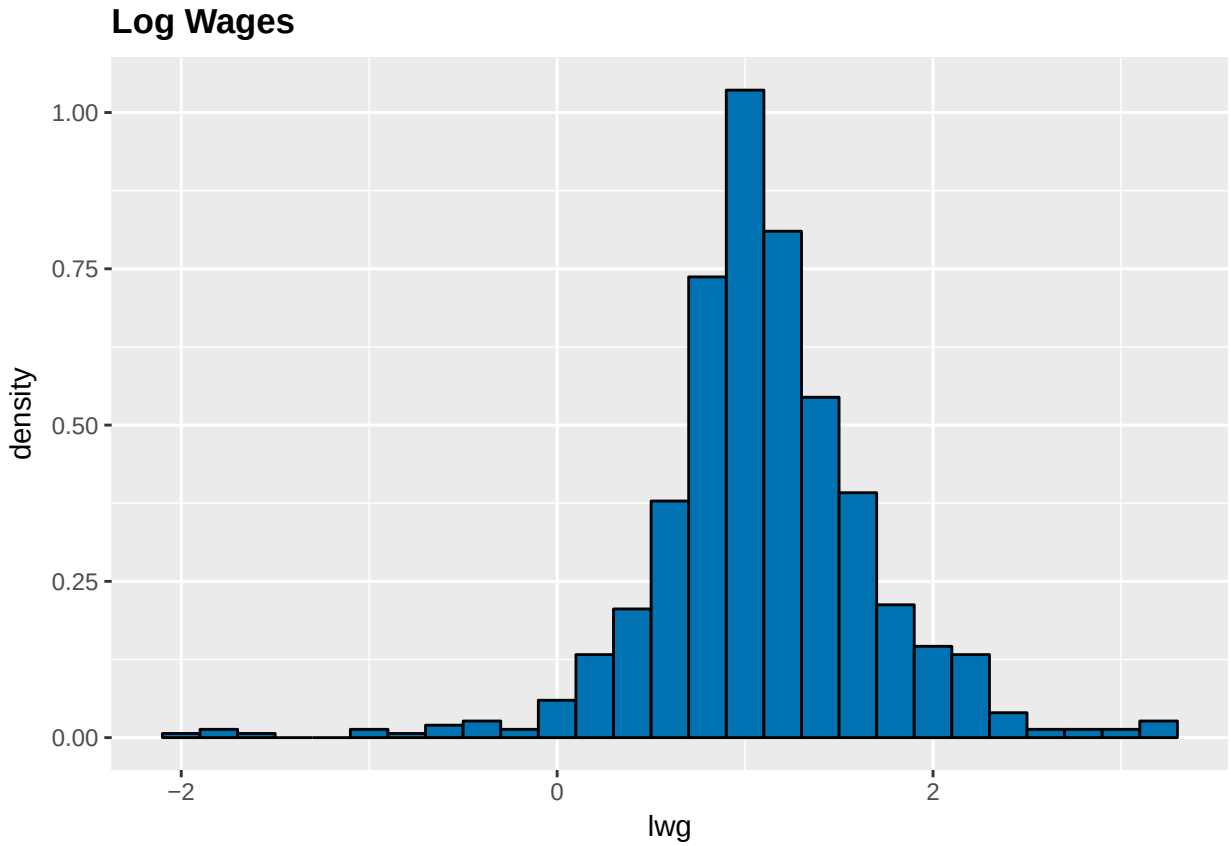
```
require(dplyr)
describe(exp(Mroz$lwg))
```

```
## exp(Mroz$lwg)
##       n missing  unique    Info    Mean     .05     .10     .25     .50
##     753       0     676       1   3.567   1.242   1.646   2.266   2.911
##     .75     .90     .95
##   4.054   5.812   7.967
##
## lowest :  0.1282  0.1616  0.1709  0.2137  0.3571
## highest: 18.2667 21.4286 22.5000 23.4667 25.0000
```

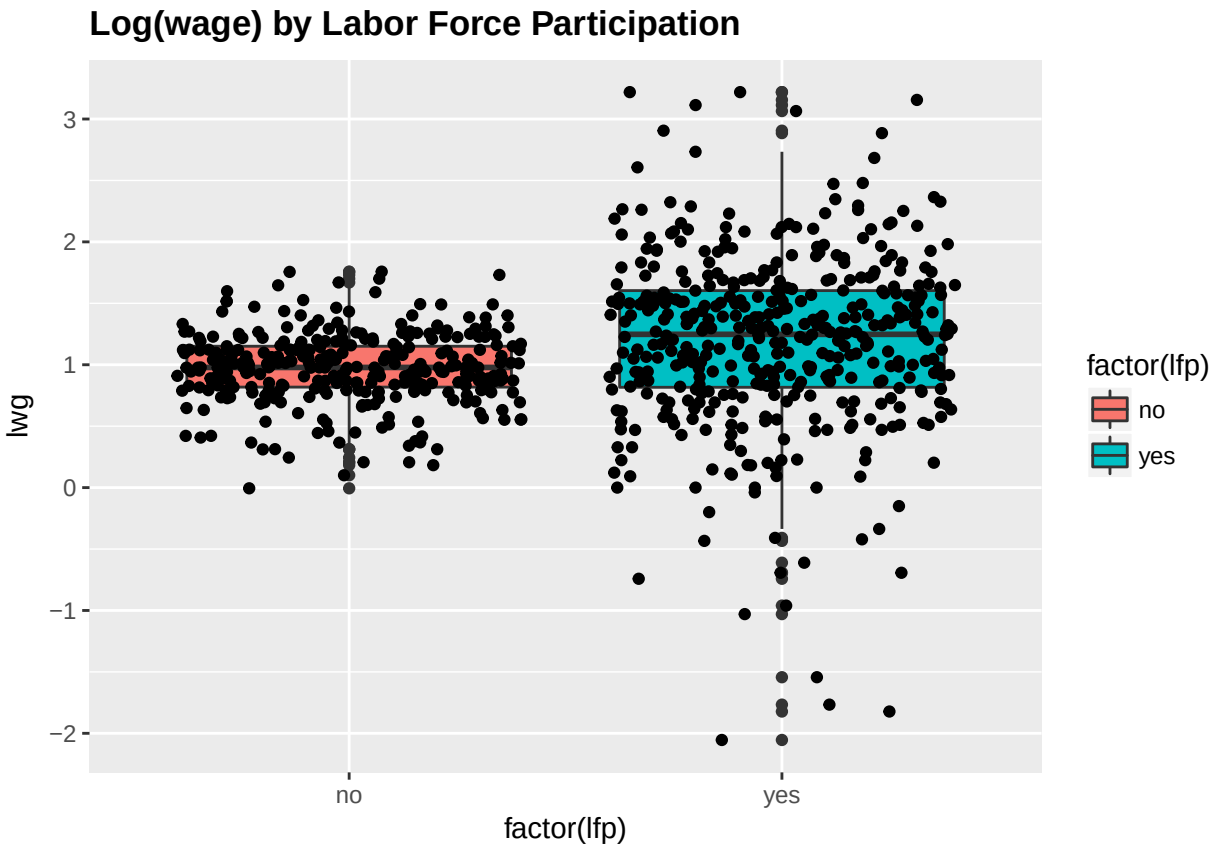
```
min(exp(Mroz$lwg))
```

```
## [1] 0.1282051
```

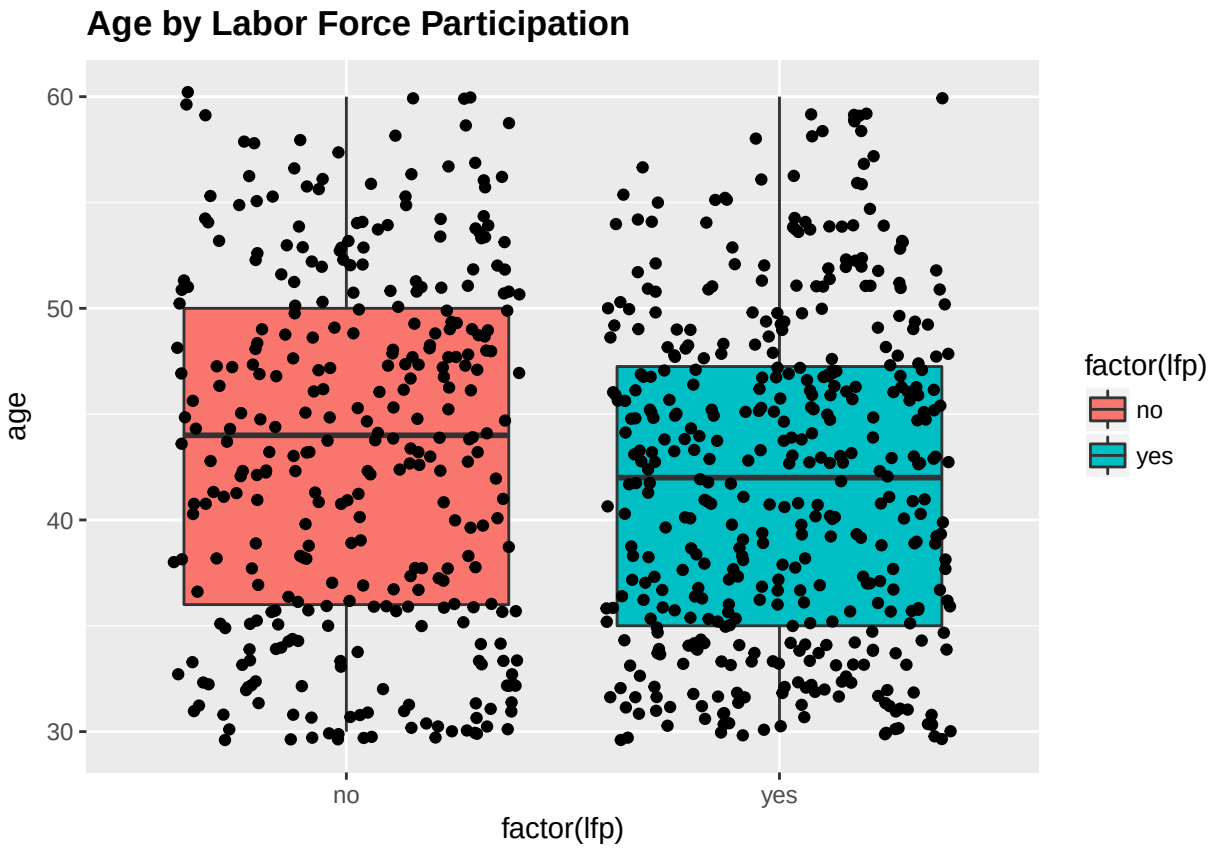
```
require(ggplot2)
# require(GGally)

# Distribution of log(wage)
ggplot(Mroz, aes(x = lwg)) + geom_histogram(aes(y = ..density..),
    binwidth = 0.2, fill = "#0072B2", colour = "black") + ggtitle("Log Wages") +
    theme(plot.title = element_text(lineheight = 1, face = "bold"))
```




```
# log(wage) by lfp
ggplot(Mroz, aes(factor(lfp), lwg)) + geom_boxplot(aes(fill = factor(lfp))) +
    geom_jitter() + ggtitle("Log(wage) by Labor Force Participation") +
    theme(plot.title = element_text(lineheight = 1, face = "bold"))
```



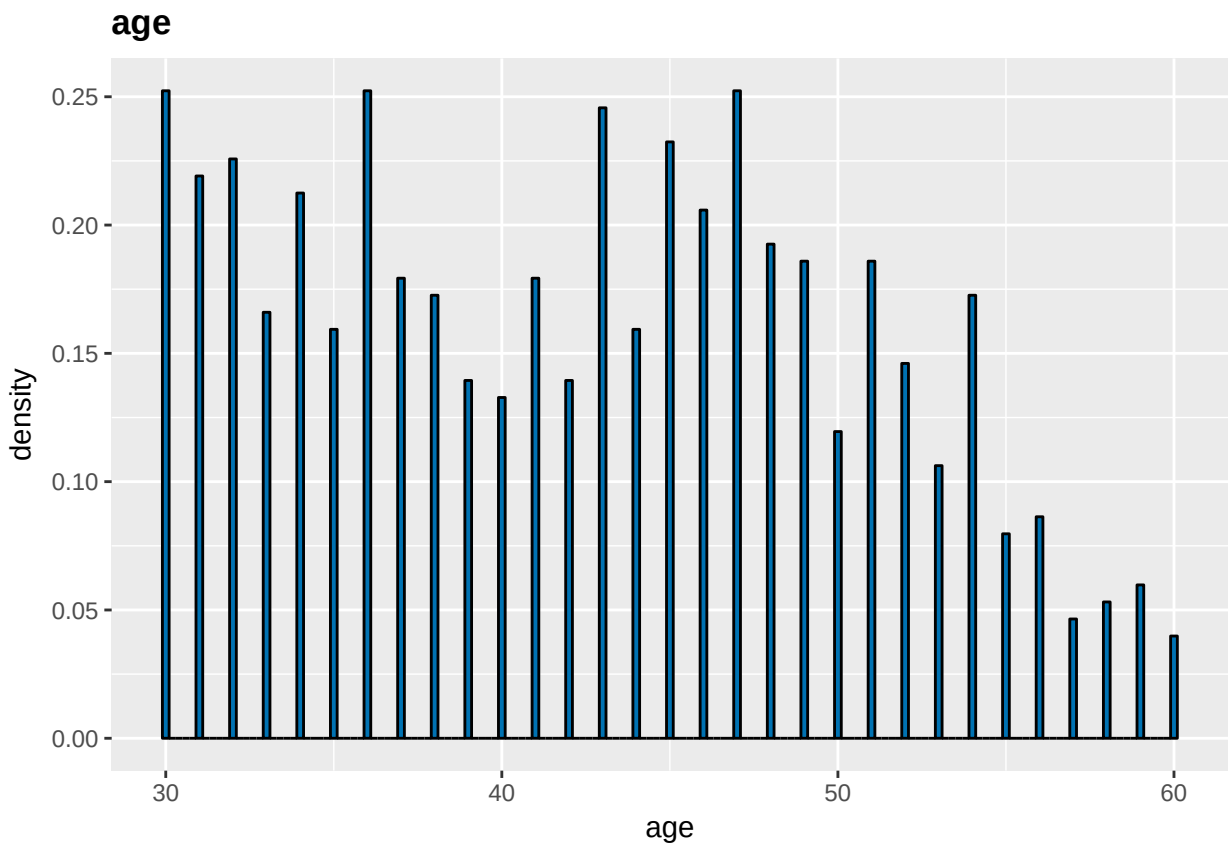

```
# age by lfp
ggplot(Mroz, aes(factor(lfp), age)) + geom_boxplot(aes(fill = factor(lfp))) +
    geom_jitter() + ggtitle("Age by Labor Force Participation") +
    theme(plot.title = element_text(lineheight = 1, face = "bold"))
```



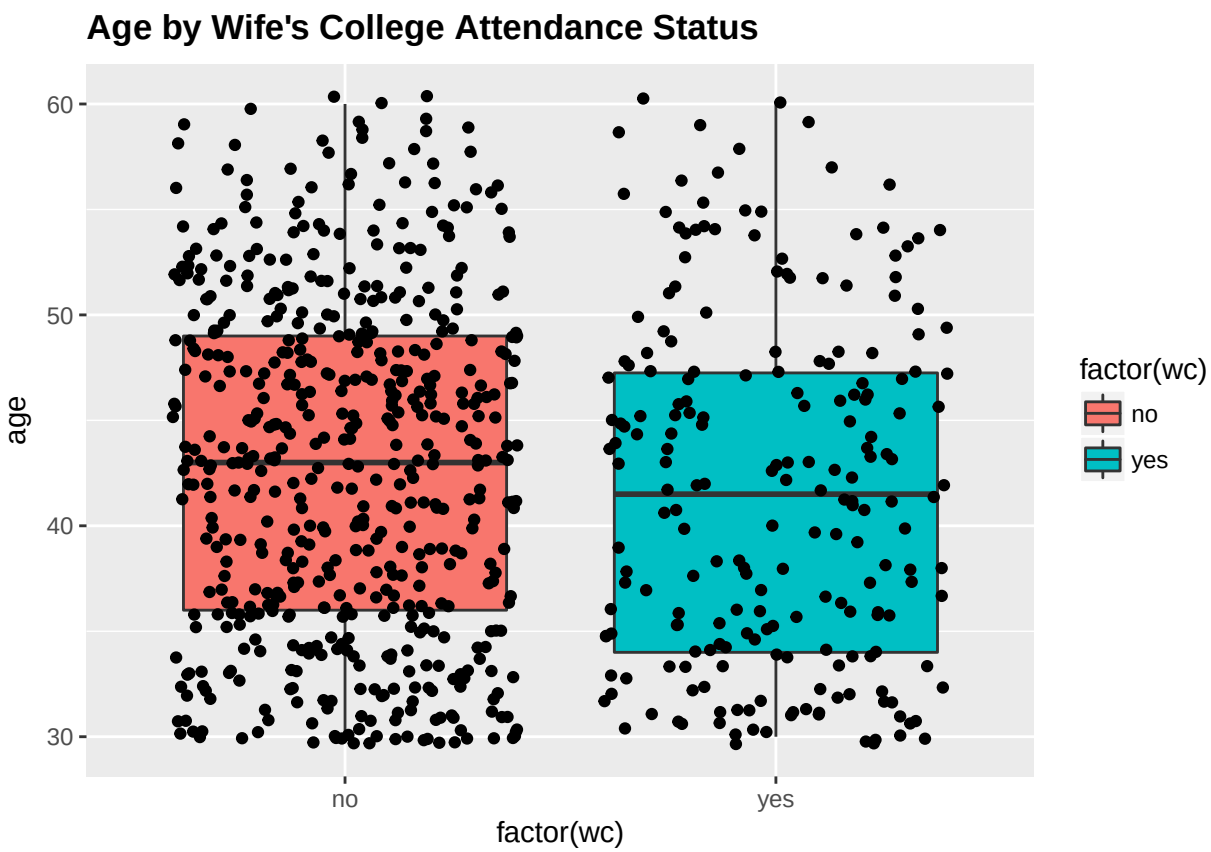

```
# Distribution of age
summary(Mroz$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   30.00   36.00   43.00   42.54   49.00   60.00
```

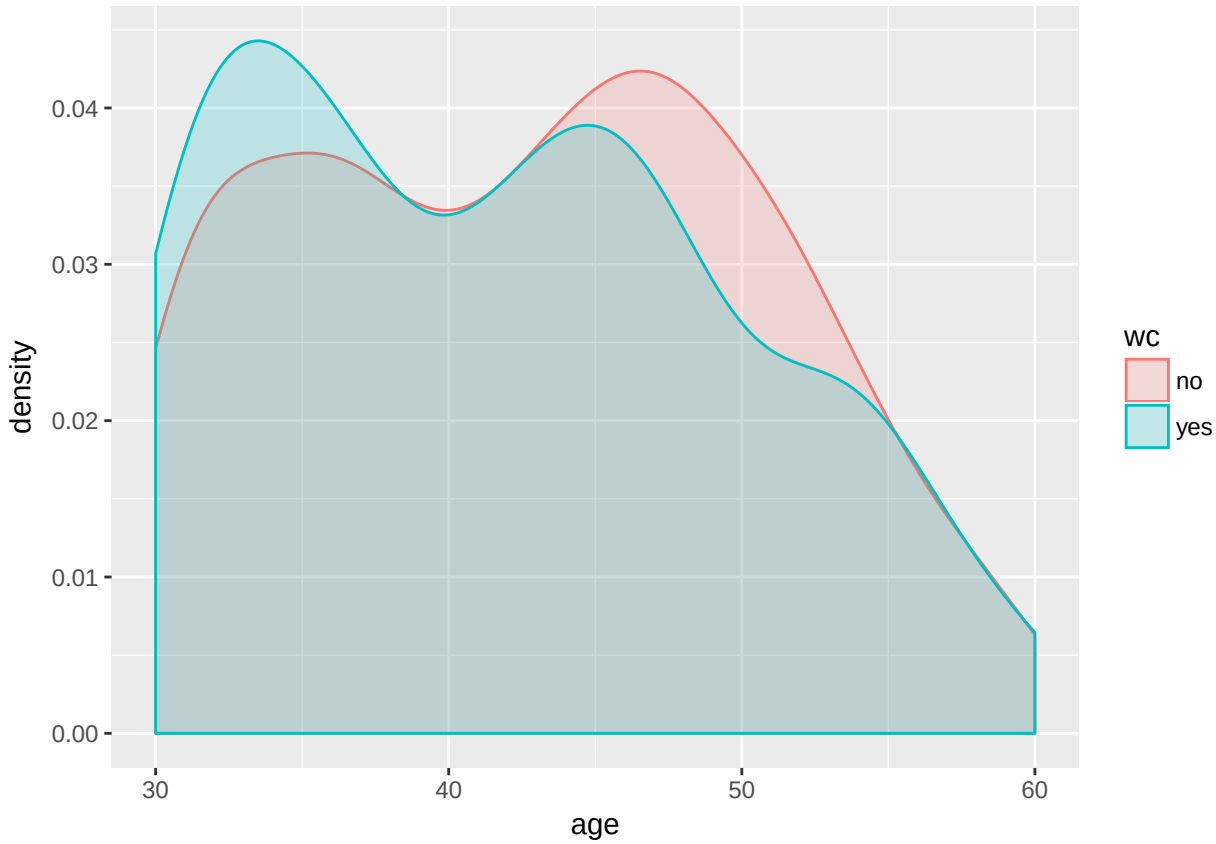
```
ggplot(Mroz, aes(x = age)) + geom_histogram(aes(y = ..density..),
    binwidth = 0.2, fill = "#0072B2", colour = "black") + ggtitle("age") +
    theme(plot.title = element_text(lineheight = 1, face = "bold"))
```



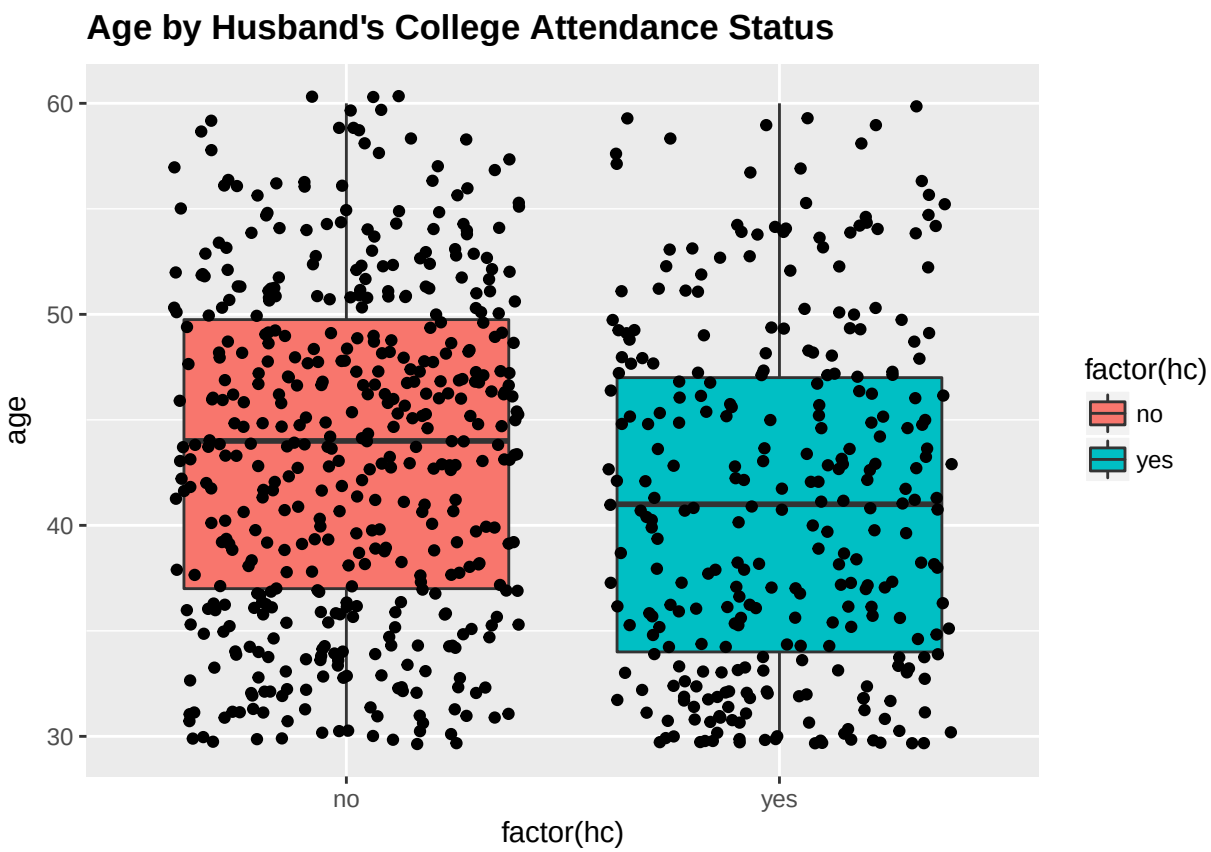

```
# Distribution of age by wc Were those who attended colleage
# tend to be younger?
ggplot(Mroz, aes(factor(wc), age)) + geom_boxplot(aes(fill = factor(wc))) +
    geom_jitter() + ggtitle("Age by Wife's College Attendance Status") +
    theme(plot.title = element_text(lineheight = 1, face = "bold"))
```



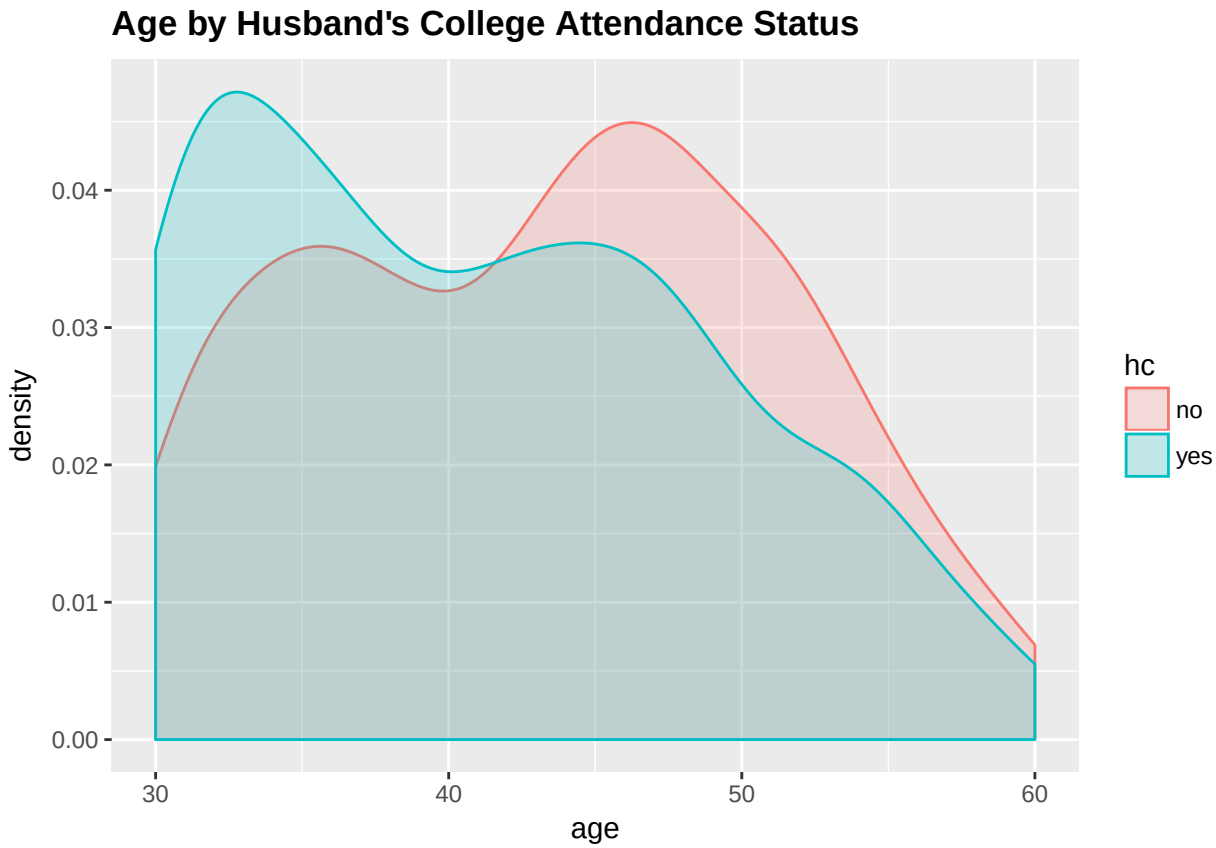

```
ggplot(Mroz, aes(age, fill = wc, colour = wc)) + geom_density(alpha = 0.2)
```



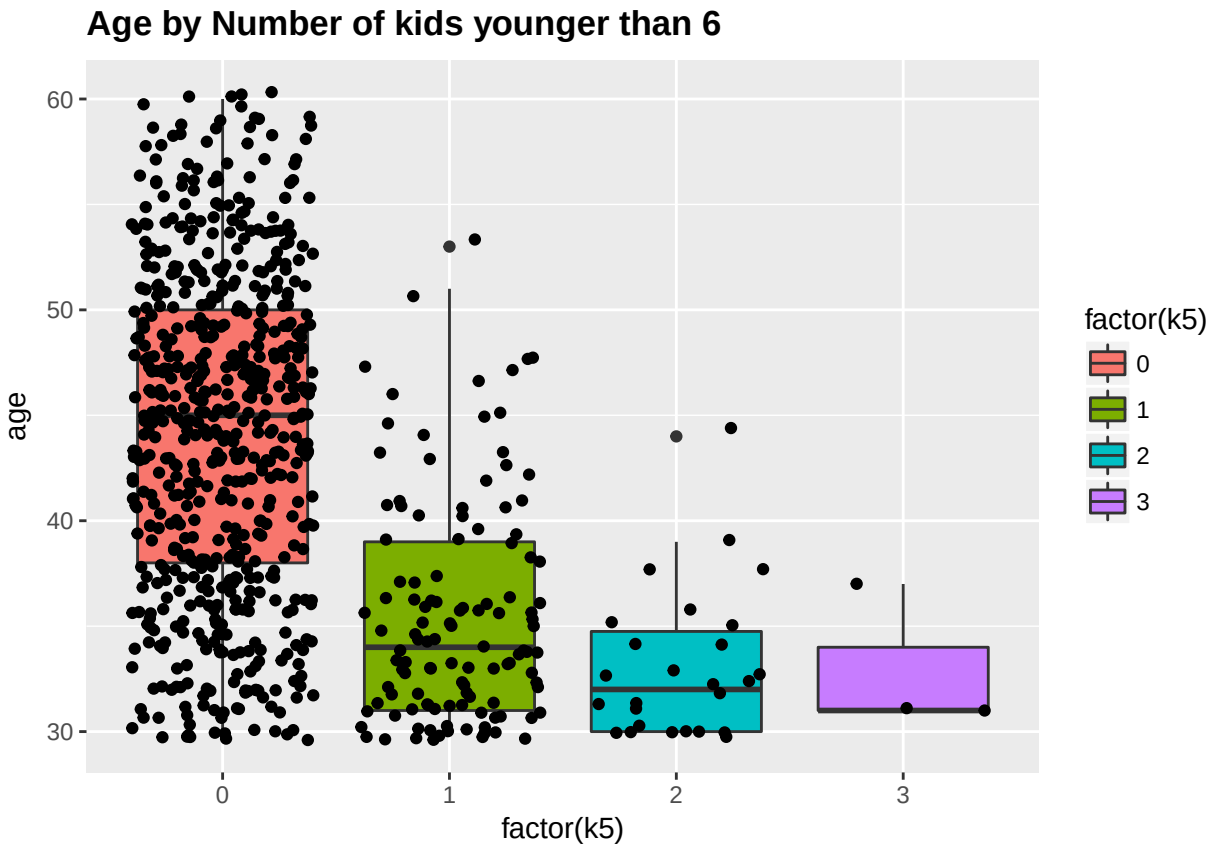

```
# Distribution of age by hc Were those whose husband attended
# colleage tend to be younger?
ggplot(Mroz, aes(factor(hc), age)) + geom_boxplot(aes(fill = factor(hc))) +
    geom_jitter() + ggtitle("Age by Husband's College Attendance Status") +
    theme(plot.title = element_text(lineheight = 1, face = "bold"))
```



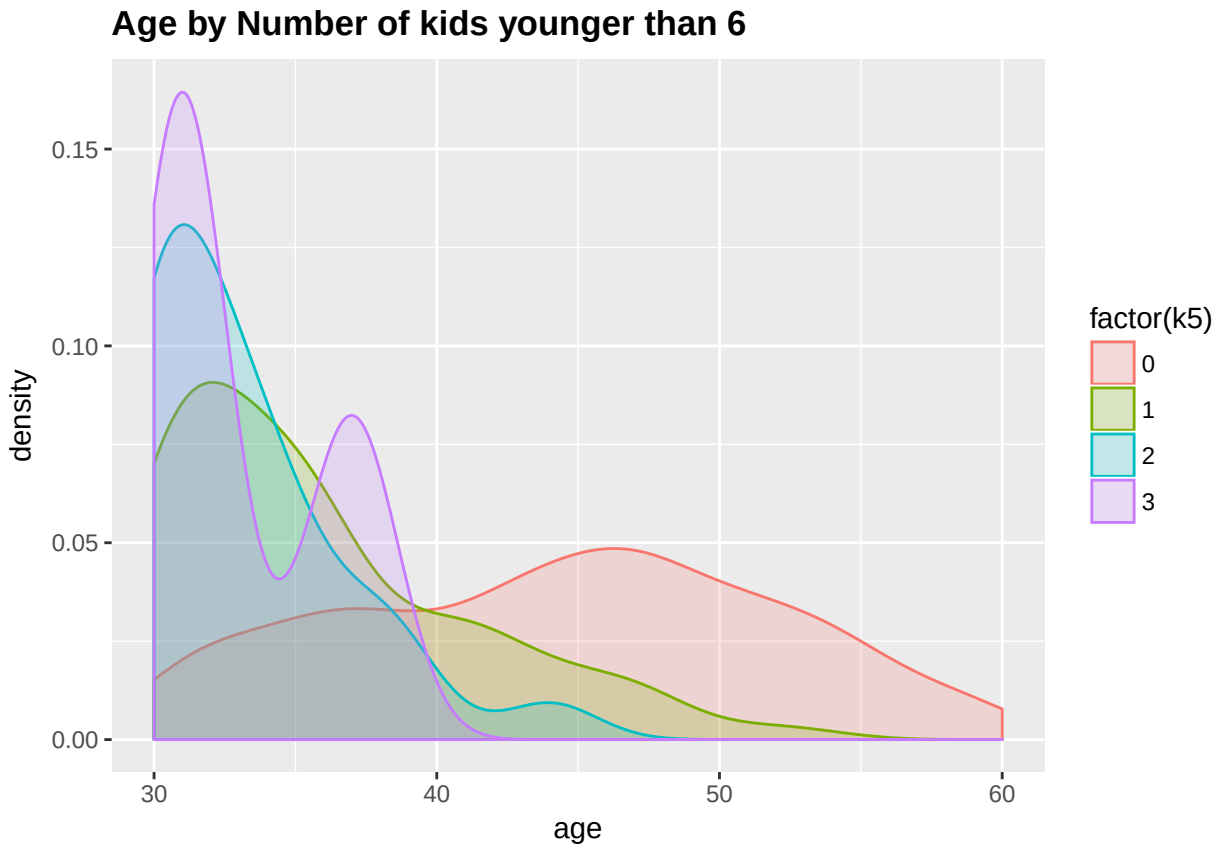

```
ggplot(Mroz, aes(age, fill = hc, colour = hc)) + geom_density(alpha = 0.2) +
    ggtitle("Age by Husband's College Attendance Status") + theme(plot.title = element_text(lineheight =
    face = "bold"))
```



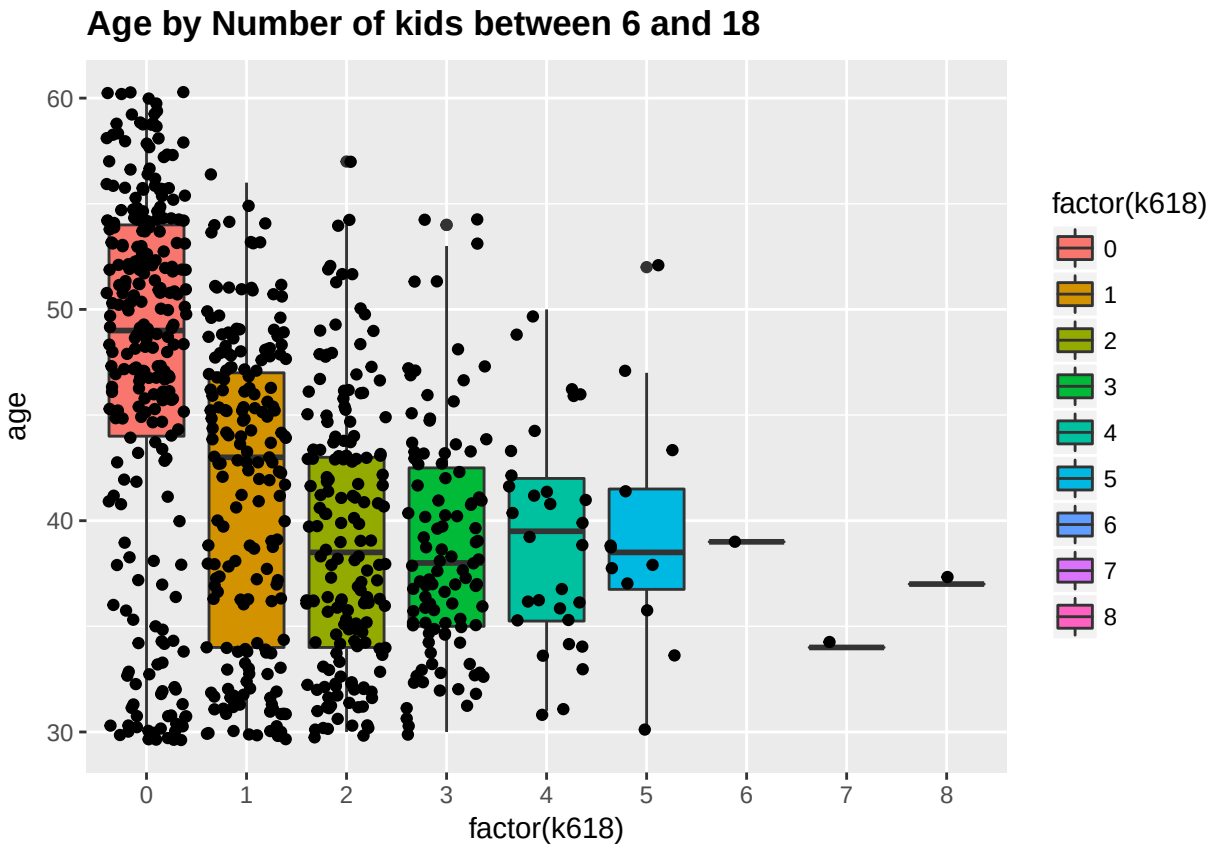

```
# Distribution of age by number kids in different age group
ggplot(Mroz, aes(factor(k5), age)) + geom_boxplot(aes(fill = factor(k5))) +
    geom_jitter() + ggtitle("Age by Number of kids younger than 6") +
    theme(plot.title = element_text(lineheight = 1, face = "bold"))
```



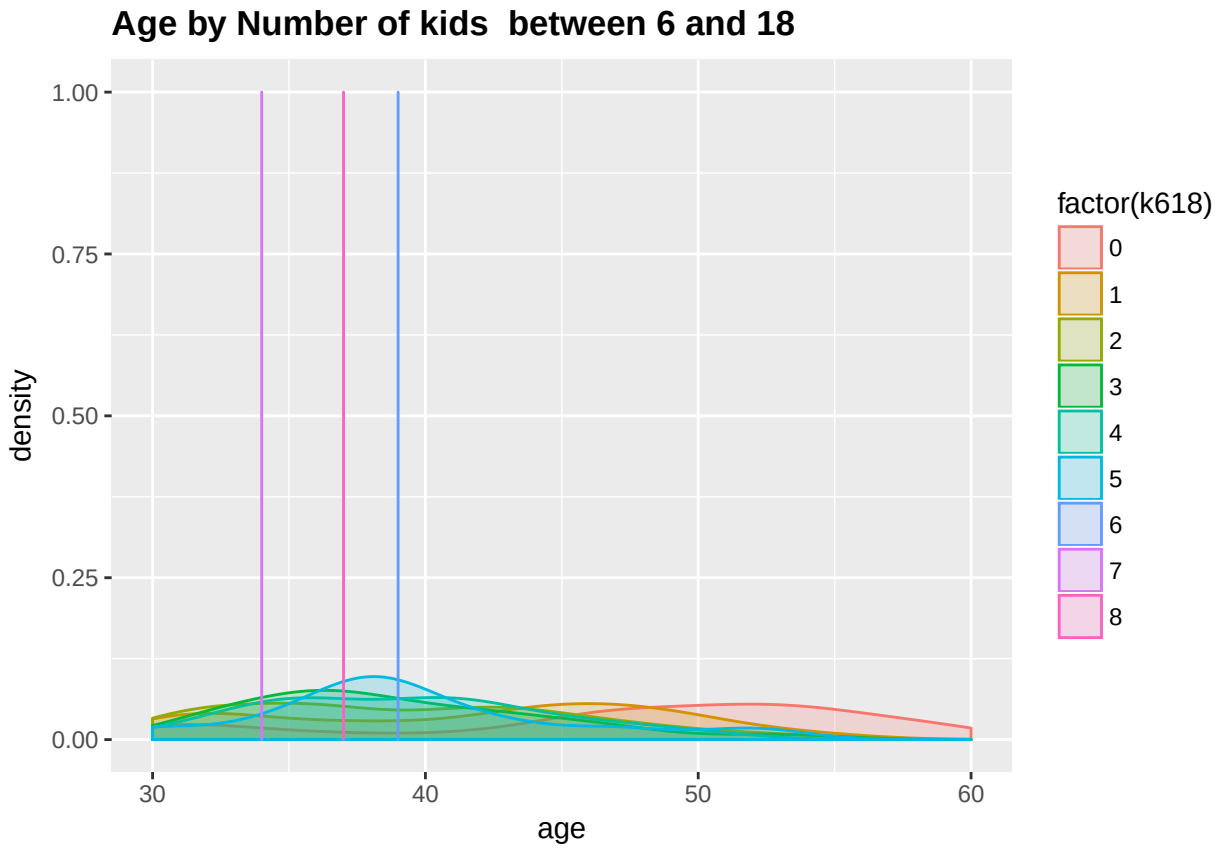

```
ggplot(Mroz, aes(age, fill = factor(k5), colour = factor(k5))) +
    geom_density(alpha = 0.2) + ggtitle("Age by Number of kids younger than 6") +
    theme(plot.title = element_text(lineheight = 1, face = "bold"))
```




```
ggplot(Mroz, aes(factor(k618), age)) + geom_boxplot(aes(fill = factor(k618))) +
    geom_jitter() + ggtitle("Age by Number of kids between 6 and 18") +
    theme(plot.title = element_text(lineheight = 1, face = "bold"))
```




```
ggplot(Mroz, aes(age, fill = factor(k618), colour = factor(k618))) +
    geom_density(alpha = 0.2) + ggtitle("Age by Number of kids  between 6 and 18") +
    theme(plot.title = element_text(lineheight = 1, face = "bold"))
```




```
# It may be easier to visualize age by first binning the
# variable
table(Mroz$k5)
```

```
##
##   0   1   2   3
## 606 118  26   3
```

```
table(Mroz$k618)
```

```
##
##   0   1   2   3   4   5   6   7   8
## 258 185 162 103  30  12   1   1   1
```

```
table(Mroz$k5, Mroz$k618)
```

```
##
##       0   1   2   3   4   5   6   7   8
##   0 229 144 121  75  26   9   0   1   1
##   1  17  35  36  24   3   3   0   0   0
##   2  11   5   5   3   1   0   1   0   0
##   3   1   1   0   1   0   0   0   0   0
```

```
xtabs(~k5 + k618, data = Mroz)
```

```
##    k618
## k5    0   1   2   3   4   5   6   7   8
##   0 229 144 121  75  26   9   0   1   1
##   1  17  35  36  24   3   3   0   0   0
##   2  11   5   5   3   1   0   1   0   0
```

```
##   3   1   1   0   1   0   0   0   0   0
```

```
table(Mroz$hc)
```

```
##
##  no yes
## 458 295
```

```
round(prop.table(table(Mroz$hc)), 2)
```

```
##
##   no  yes
## 0.61 0.39
```

```
table(Mroz$wc)
```

```
##
##  no yes
## 541 212
```

```
round(prop.table(table(Mroz$wc)), 2)
```

```
##
##   no  yes
## 0.72 0.28
```

```
xtabs(~hc + wc, data = Mroz)
```

```
##      wc
## hc     no yes
##   no  417  41
##   yes 124 171
```

```
round(prop.table(xtabs(~hc + wc, data = Mroz)), 2)
```

```
##      wc
## hc      no  yes
##   no  0.55 0.05
##   yes 0.16 0.23
```

A quick revisit of some of the concepts from lecture 1

** Test of independence: Chi-squared test TO BE HERE**

Despite the EDA conducted above suggest interaction of variables and creation of some new variables, I following the specification used in Mroz (1987)'s paper. I want to leave the interaction of variables and creation of some new variables as take-home exerciase and have you presented in the next live session when we cover specification.

Estimate a linear regression model, conduct model diagnostic, test model assumption, and interpret model results

```
mroz.lm <- lm(as.numeric(lfp) ~ k5 + k618 + age + wc + hc + lwg +
    inc, data = Mroz)
summary(mroz.lm)
```

```
##
```

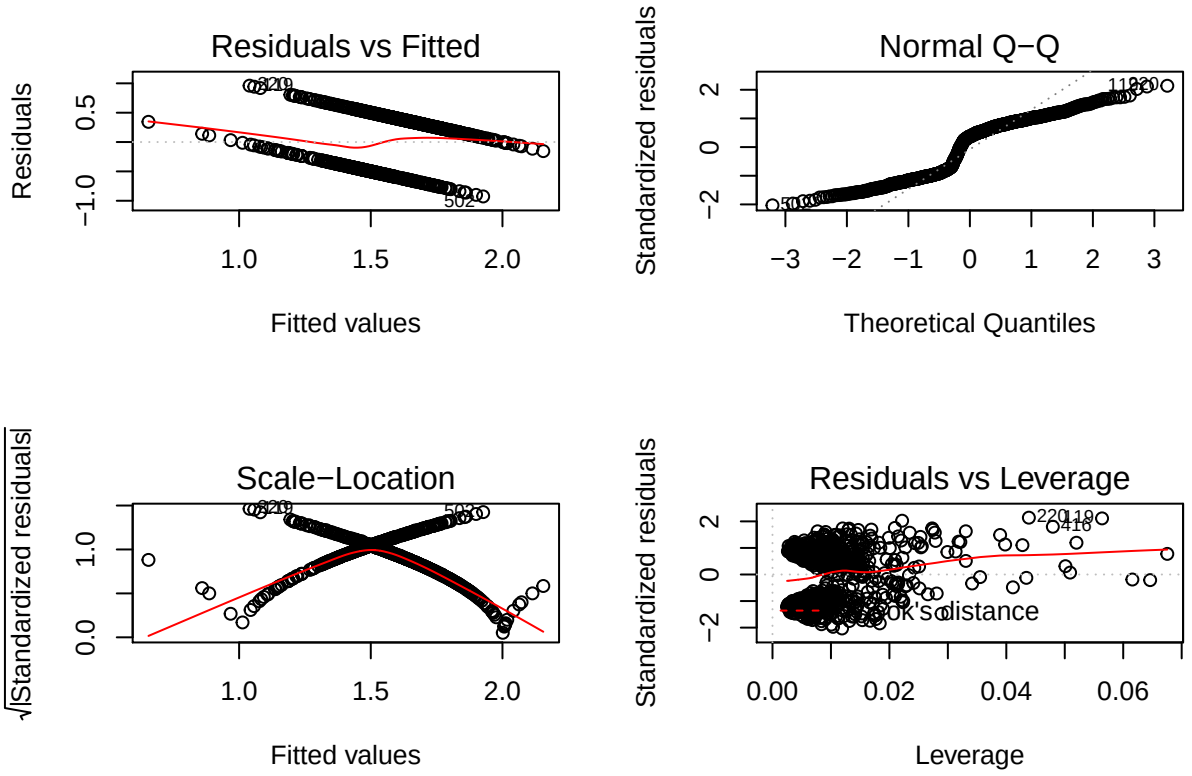
```
## Call:
## lm(formula = as.numeric(lfp) ~ k5 + k618 + age + wc + hc + lwg +
##     inc, data = Mroz)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9268 -0.4632  0.1684  0.3906  0.9602
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.143548   0.127053  16.871  < 2e-16 ***
## k5          -0.294836   0.035903  -8.212 9.58e-16 ***
## k618        -0.011215   0.013963  -0.803 0.422109
## age         -0.012741   0.002538  -5.021 6.45e-07 ***
## wcyes        0.163679   0.045828   3.572 0.000378 ***
## hcyes        0.018951   0.042533   0.446 0.656044
## lwg          0.122740   0.030191   4.065 5.31e-05 ***
## inc         -0.006760   0.001571  -4.304 1.90e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.459 on 745 degrees of freedom
## Multiple R-squared:  0.1503, Adjusted R-squared:  0.1423
## F-statistic: 18.83 on 7 and 745 DF,  p-value: < 2.2e-16
```

Model diagnotic

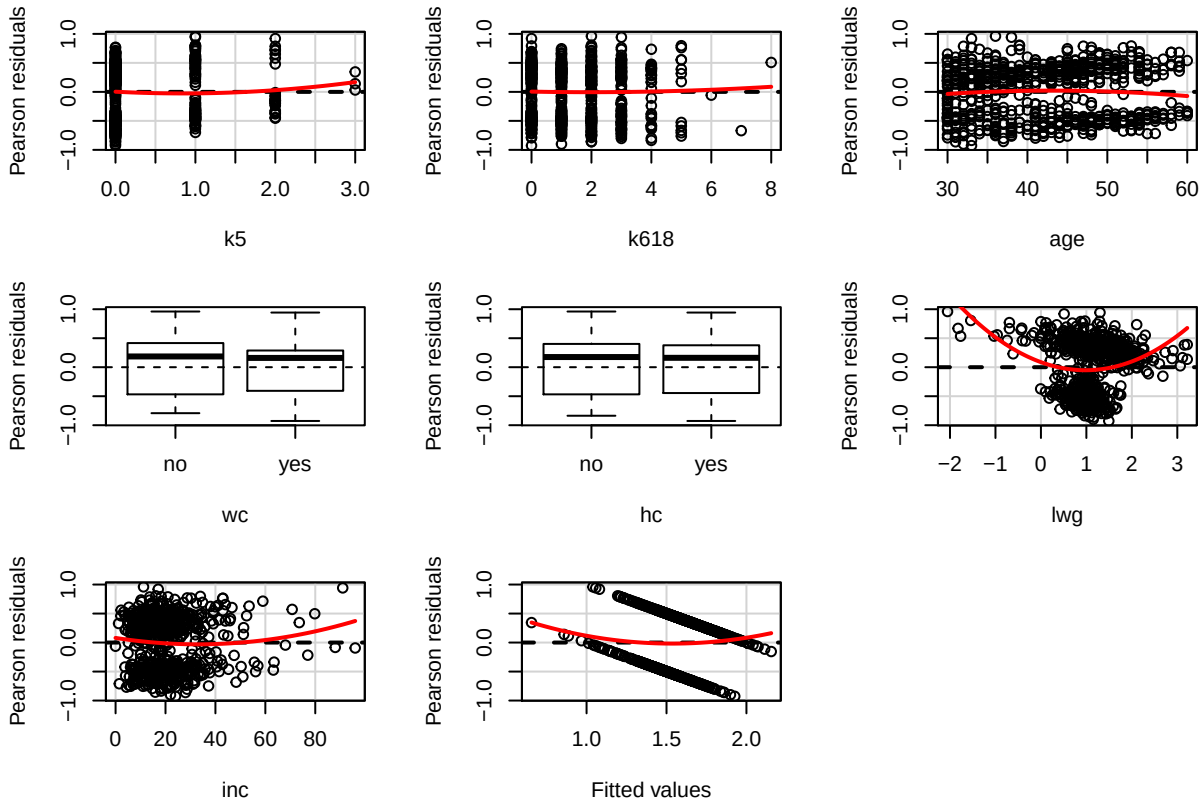
Exercise (5 minutes): 1. Interpret the diagnostic plots. 2. Discuss the impact of using linear probability model on fitted values. Write more codes to aid your discussion where needed.

First and foremost, the ploto f the Pearson residuals against fitted values do not appear to be random at all; it shows a very strong patterns. More importantly, most of the fitted value goes beyond 1.

```
par(mfrow = c(2, 2))
plot(mroz.lm)
```



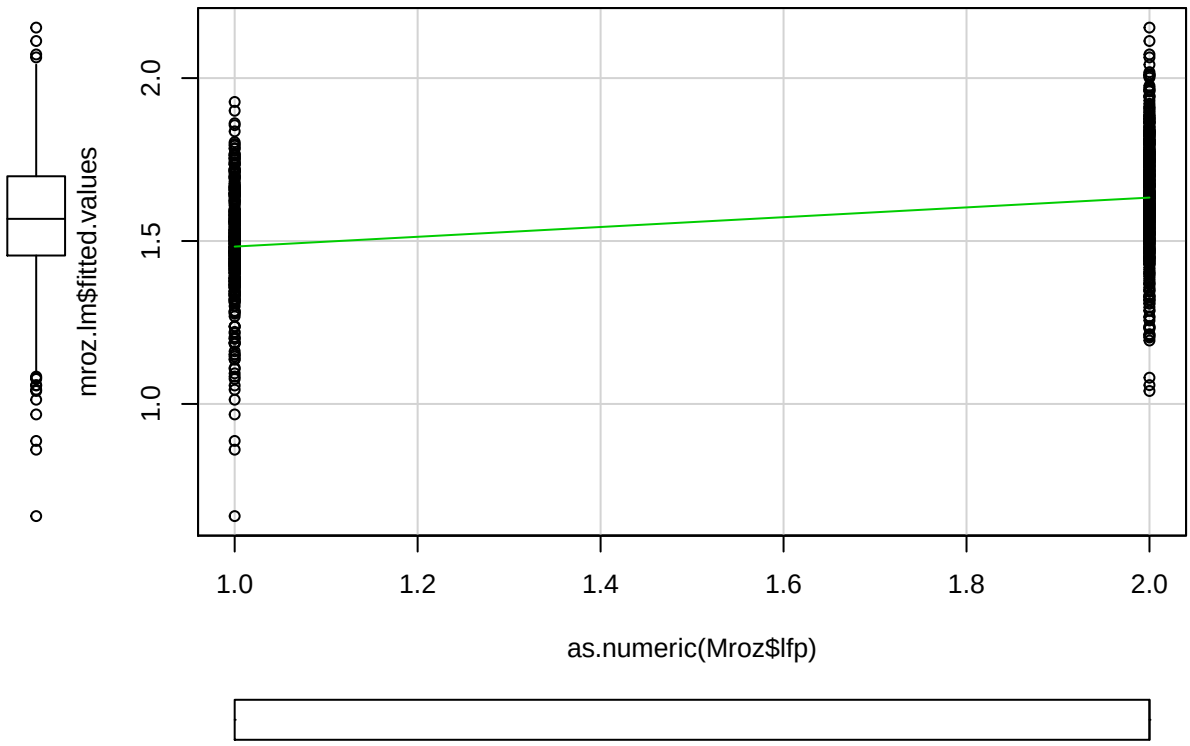

```
require(car)
par(mfrow = c(1, 1))
residualPlots(mroz.lm)
```



```
##            Test stat Pr(>|t|)
## k5             0.969    0.333
## k618           0.384    0.701
## age           -1.347    0.178
## wc                NA       NA
## hc                NA       NA
## lwg            7.697    0.000
## inc            1.970    0.049
## Tukey test     2.035    0.042
```

```
# Another way to diagnose linearity assumptions
scatterplot(as.numeric(Mroz$lfp), mroz.lm$fitted.values)
```

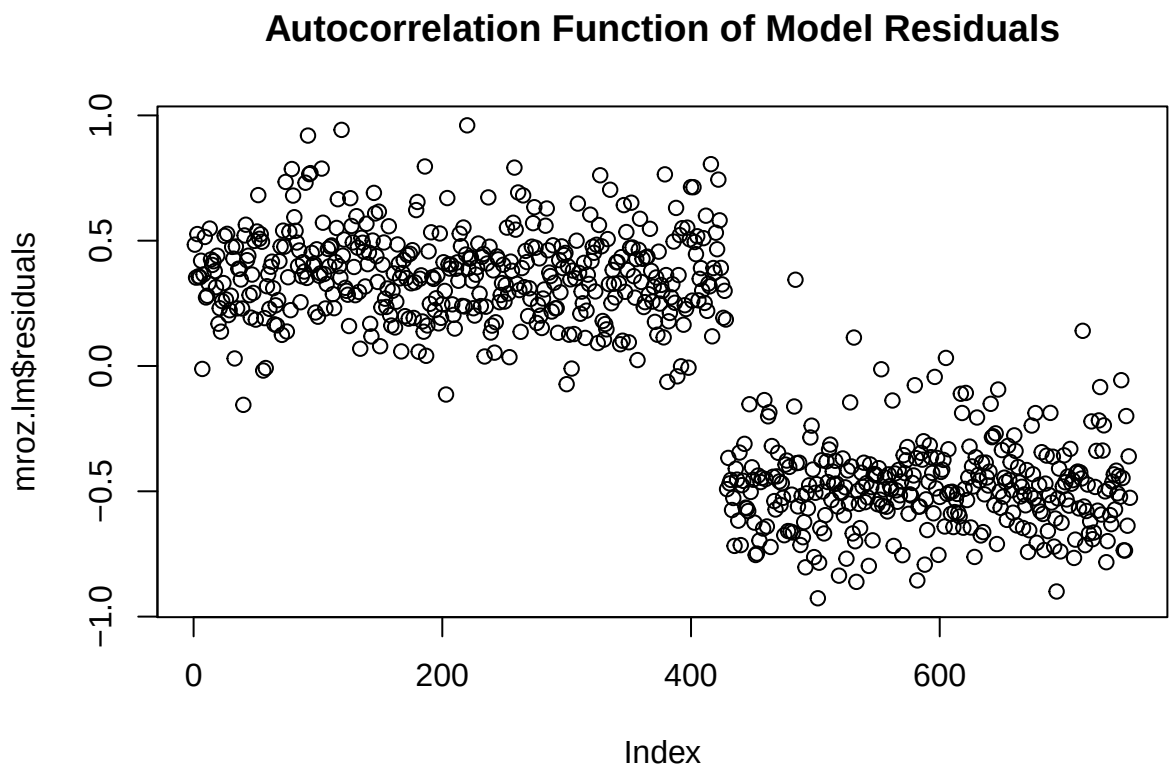
```
## Warning in smoother(.x, .y, col = col[2], log.x = logged("x"), log.y =
## logged("y"), : could not fit smooth
```



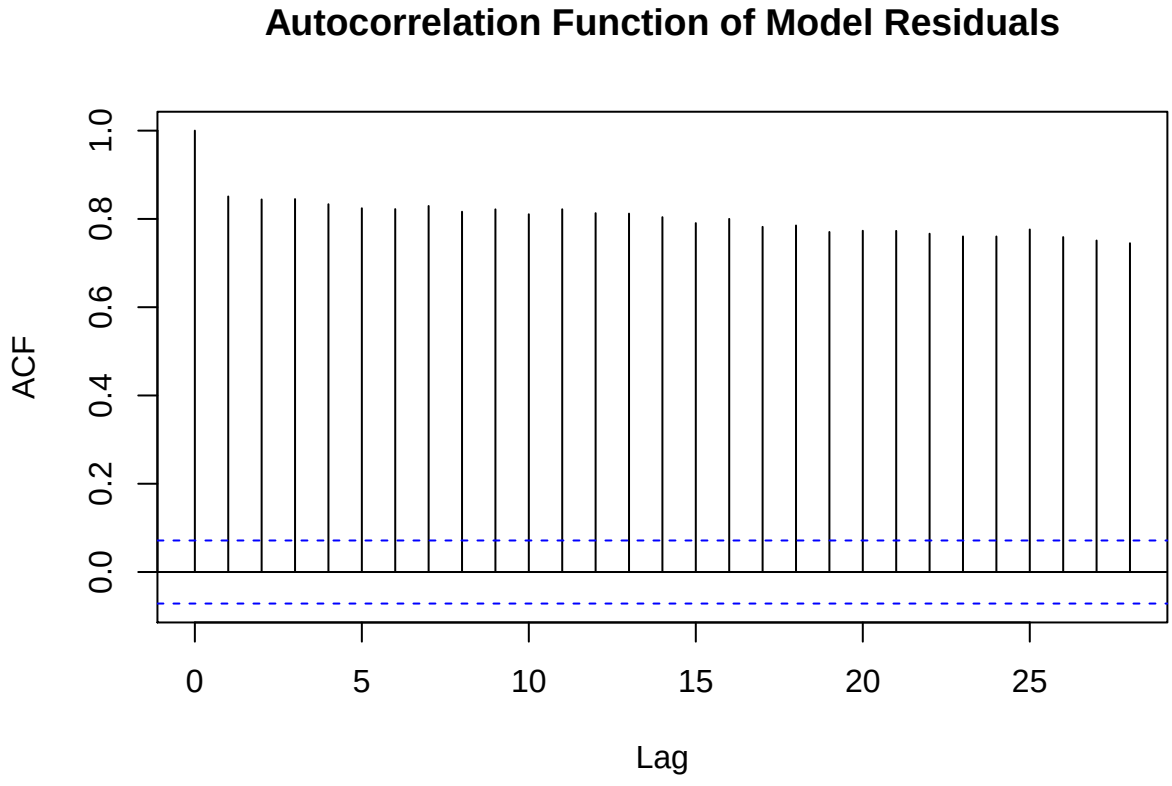

```
# Note that I didn't pay much attention to outliers and
# influential observations in this specific example

# Evaluate Nonlinearity component + residual plot
# crPlots(mroz.lm)

# Ceres plots ceresPlots(mroz.lm)

# str(mroz.lm)
summary(mroz.lm$fitted.values)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.6558  1.4560  1.5680  1.5680  1.6990  2.1550
```

```
par(mfrow = c(1, 1))
plot(mroz.lm$residuals, main = "Autocorrelation Function of Model Residuals")
```

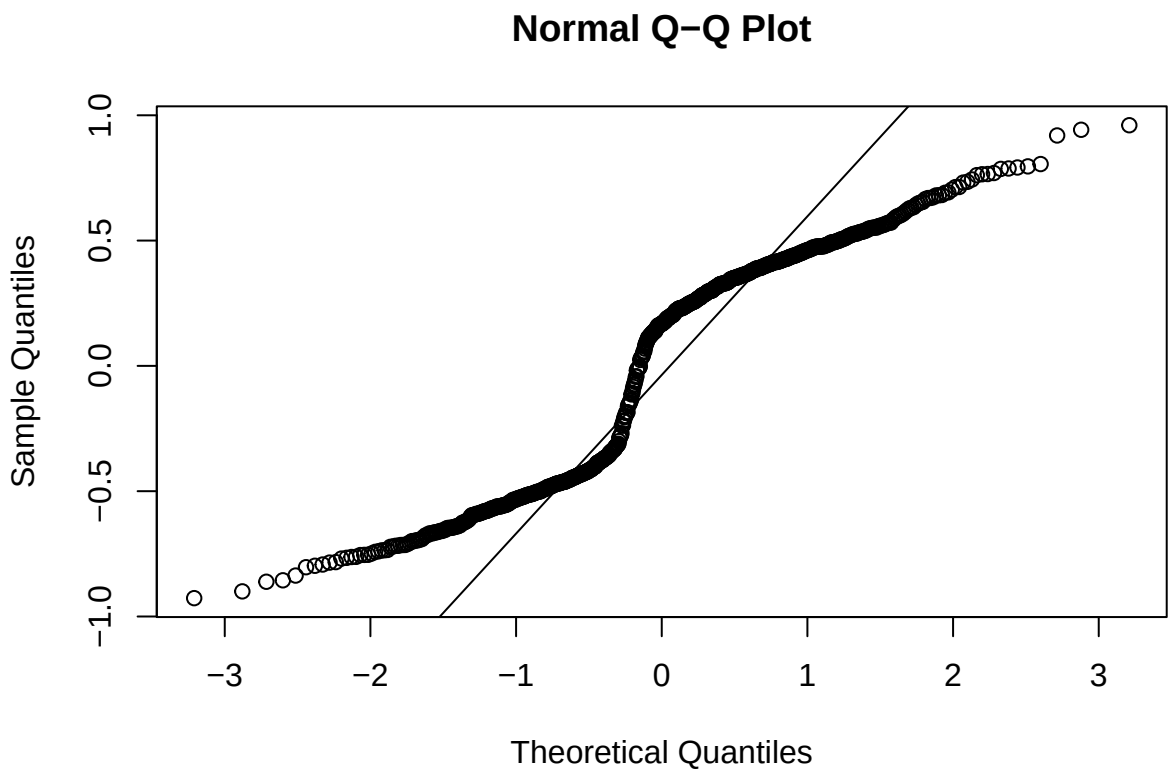



```
acf(mroz.lm$residuals, main = "Autocorrelation Function of Model Residuals")
```



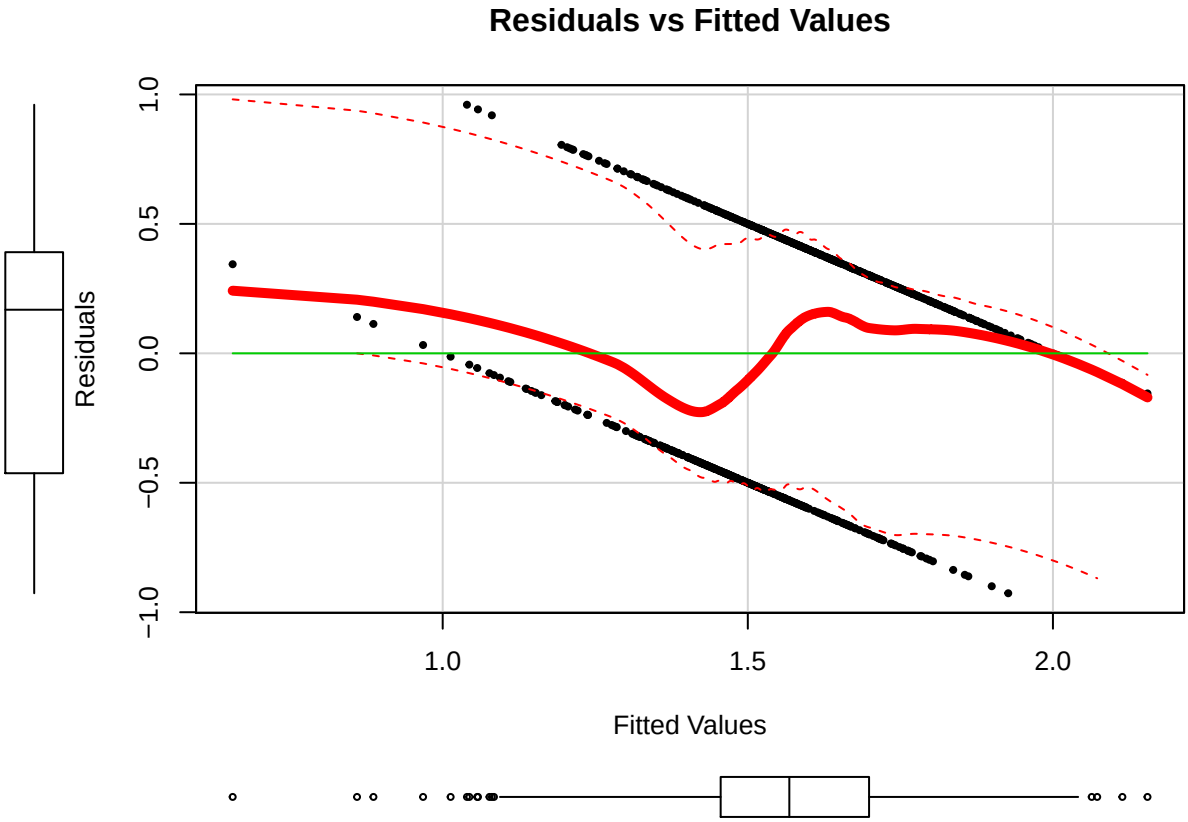

```
hist(mroz.lm$residuals)
```


```
qqnorm(mroz.lm$residuals)
qqline(mroz.lm$residuals)
```




```
scatterplot(mroz.lm$fitted.values, mroz.lm$residuals, smoother = loessLine,
    cex = 0.5, pch = 19, smoother.args = list(lty = 1, lwd = 5),
```

```
    main = "Residuals vs Fitted Values", xlab = "Fitted Values",
    ylab = "Residuals")
```



Test CLM model assumptions

Exercise (5 minutes): Interpret the results of each of these tests.

```
# Test of Normality
shapiro.test(mroz.lm$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mroz.lm$residuals
## W = 0.91081, p-value < 2.2e-16
```

```
# Heteroskedasticity: Non-constant Variance Test
require(car)
ncvTest(mroz.lm)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 9.562012    Df = 1     p = 0.001986453
```

```
# http://math.furman.edu/~dcs/courses/math47/R/library/lmtest/html/bptest.html
require(lmtest)
```

```
## Loading required package: lmtest
```

```
## Loading required package: zoo
```

```
## Warning: package 'zoo' was built under R version 3.2.5
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
bptest(mroz.lm)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mroz.lm
## BP = 97.603, df = 7, p-value < 2.2e-16
```

```
# Test of Independence
durbinWatsonTest(mroz.lm)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.8508422     0.2950591       0
##  Alternative hypothesis: rho != 0
```

```
# Global test of model assumptions
# https://cran.r-project.org/web/packages/gvlma/index.html
# install.packages('gvlma')
require(gvlma)
```

```
## Loading required package: gvlma
```

```
gv.mroz.lm <- gvlma(mroz.lm)
summary(gv.mroz.lm)
```

```
##
## Call:
## lm(formula = as.numeric(lfp) ~ k5 + k618 + age + wc + hc + lwg +
##     inc, data = Mroz)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9268 -0.4632  0.1684  0.3906  0.9602
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.143548   0.127053  16.871  < 2e-16 ***
## k5          -0.294836   0.035903  -8.212 9.58e-16 ***
## k618        -0.011215   0.013963  -0.803 0.422109
## age         -0.012741   0.002538  -5.021 6.45e-07 ***
## wcyes        0.163679   0.045828   3.572 0.000378 ***
## hcyes        0.018951   0.042533   0.446 0.656044
## lwg          0.122740   0.030191   4.065 5.31e-05 ***
## inc         -0.006760   0.001571  -4.304 1.90e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.459 on 745 degrees of freedom
## Multiple R-squared:  0.1503, Adjusted R-squared:  0.1423
```

```
## F-statistic: 18.83 on 7 and 745 DF,  p-value: < 2.2e-16
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = mroz.lm)
##
##                     Value   p-value                   Decision
## Global Stat        85.271 0.000e+00 Assumptions NOT satisfied!
## Skewness            4.298 3.816e-02 Assumptions NOT satisfied!
## Kurtosis           63.409 1.665e-15 Assumptions NOT satisfied!
## Link Function       4.168 4.120e-02 Assumptions NOT satisfied!
## Heteroscedasticity 13.395 2.522e-04 Assumptions NOT satisfied!
```

Interpret model results

Exercise (10 minutes): Interpret the model results. As an example, an increase in 1 child with age less than 6 decreased probability of LFP by almost 30%, holding other variables in the model constant. *Does this impact make sense to you? Please explain.*

Estimate a binary logistic regression

```
mroz.glm <- glm(lfp ~ k5 + k618 + age + wc + hc + lwg + inc,
    family = binomial, data = Mroz)
summary(mroz.glm)
```

```
##
## Call:
## glm(formula = lfp ~ k5 + k618 + age + wc + hc + lwg + inc, family = binomial,
##     data = Mroz)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1062  -1.0900   0.5978   0.9709   2.1893
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.182140   0.644375   4.938 7.88e-07 ***
## k5          -1.462913   0.197001  -7.426 1.12e-13 ***
## k618        -0.064571   0.068001  -0.950 0.342337
## age         -0.062871   0.012783  -4.918 8.73e-07 ***
## wcyes        0.807274   0.229980   3.510 0.000448 ***
## hcyes        0.111734   0.206040   0.542 0.587618
## lwg          0.604693   0.150818   4.009 6.09e-05 ***
## inc         -0.034446   0.008208  -4.196 2.71e-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:  905.27  on 745  degrees of freedom
## AIC: 921.27
##
## Number of Fisher Scoring iterations: 4
```

```
round(exp(cbind(Estimate = coef(mroz.glm), confint(mroz.glm))),
    2)
```

```
## Waiting for profiling to be done...
```

```
##             Estimate 2.5 % 97.5 %
## (Intercept)    24.10  6.94  87.03
## k5              0.23  0.16   0.34
## k618            0.94  0.82   1.07
## age             0.94  0.92   0.96
## wcyes           2.24  1.43   3.54
## hcyes           1.12  0.75   1.68
## lwg             1.83  1.37   2.48
## inc             0.97  0.95   0.98
```

Interpretation of model results

Exercise (Total: 20 minutes, including 10 minutes in Breakout room): Interpret everything in the summary of the model results. Interpret both the estimated coefficients in the original model result summary as well as their exponentiated versoin. Why do we exponentiate the coefficients? Interpret the effect (in terms of odds rations) of increasing k5 by 1-unit. Interpret the effect (in terms of odds rations) of increasing age by 5-units. Does it matter if the increase is from 30 to 35 or from 45 to 50?

Visualize the effect of family income on Female LFP

Exercise (whole class 10 minutes): Discuss the effect of family income on Female LFP

```
round(exp(cbind(Estimate = coef(mroz.glm), confint(mroz.glm))),
    2)
```

```
## Waiting for profiling to be done...
```

```
##             Estimate 2.5 % 97.5 %
## (Intercept)    24.10  6.94  87.03
## k5              0.23  0.16   0.34
## k618            0.94  0.82   1.07
## age             0.94  0.92   0.96
## wcyes           2.24  1.43   3.54
## hcyes           1.12  0.75   1.68
## lwg             1.83  1.37   2.48
## inc             0.97  0.95   0.98
```

```
summary(Mroz)
```

```
##   lfp            k5              k618            age          wc
##  no :325   Min.   :0.0000   Min.   :0.000   Min.   :30.00   no :541
##  yes:428   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:36.00   yes:212
##            Median :0.0000   Median :1.000   Median :43.00
##            Mean   :0.2377   Mean   :1.353   Mean   :42.54
```

```
##            3rd Qu.:0.0000   3rd Qu.:2.000   3rd Qu.:49.00
##            Max.   :3.0000   Max.   :8.000   Max.   :60.00
##    hc           lwg               inc
##  no :458   Min.   :-2.0541   Min.   :-0.029
##  yes:295   1st Qu.: 0.8181   1st Qu.:13.025
##            Median : 1.0684   Median :17.700
##            Mean   : 1.0971   Mean   :20.129
##            3rd Qu.: 1.3997   3rd Qu.:24.466
##            Max.   : 3.2189   Max.   :96.000
```

```
mroz.glm$coefficients
```

```
## (Intercept)          k5        k618         age       wcyes       hcyes
##  3.18214046 -1.46291304 -0.06457068 -0.06287055  0.80727378  0.11173357
##         lwg         inc
##  0.60469312 -0.03444643
```

```
str(mroz.glm$coefficients)
```

```
##  Named num [1:8] 3.1821 -1.4629 -0.0646 -0.0629 0.8073 ...
##  - attr(*, "names")= chr [1:8] "(Intercept)" "k5" "k618" "age" ...
```

```
coef <- mroz.glm$coefficients
coef[1]
```

```
## (Intercept)
##     3.18214
```

```
min(Mroz$inc)
```

```
## [1] -0.029
```

```
# Effect of income on LFP for a family with no kid, wife was
# 40 years old, both wife and husband attended college, and
# wife's estimated wage rate was 1.07

rm(x)
```

```
## Warning in rm(x): object 'x' not found
```

```
xx = c(1, 0, 0, 40, 1, 1, 1.07)
length(coef)
```

```
## [1] 8
```

```
length(xx)
```

```
## [1] 7
```

```
z = coef[1] * xx[1] + coef[2] * xx[2] + coef[3] * xx[3] + coef[3] *
    xx[3] + coef[4] * xx[4] + coef[5] * xx[5] + coef[6] * xx[6] +
    coef[7] * xx[7]
z
```

```
## (Intercept)
##    2.233347
```

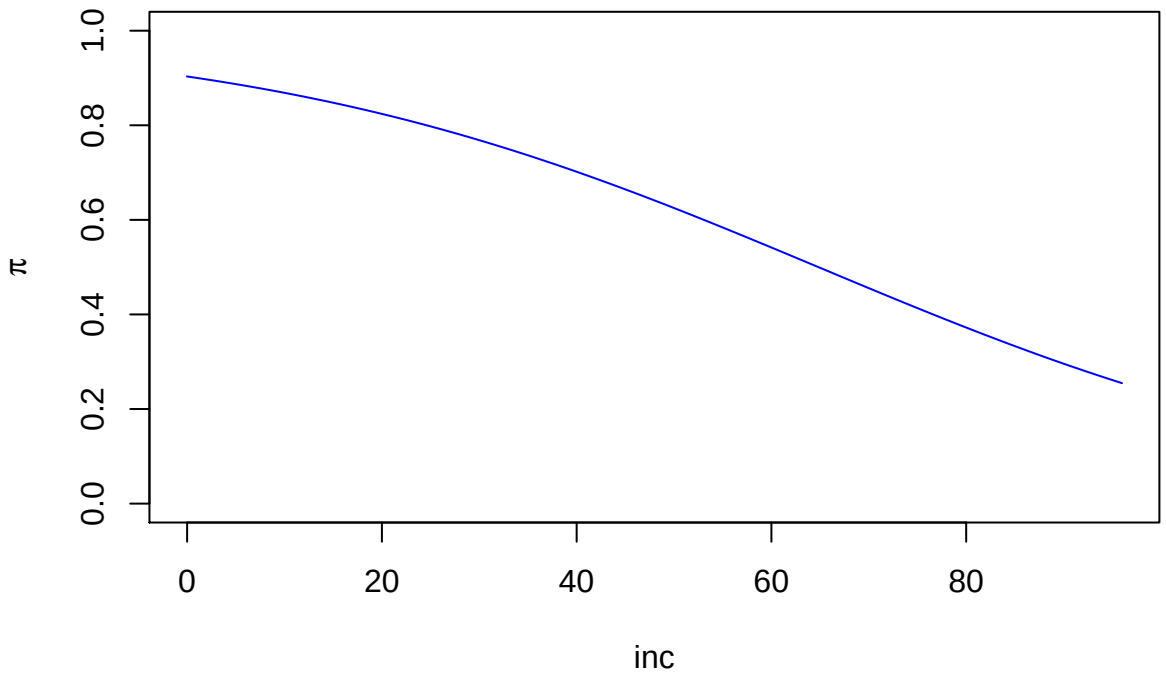
```
x <- Mroz$inc
coef[8]
```

```
##         inc
```

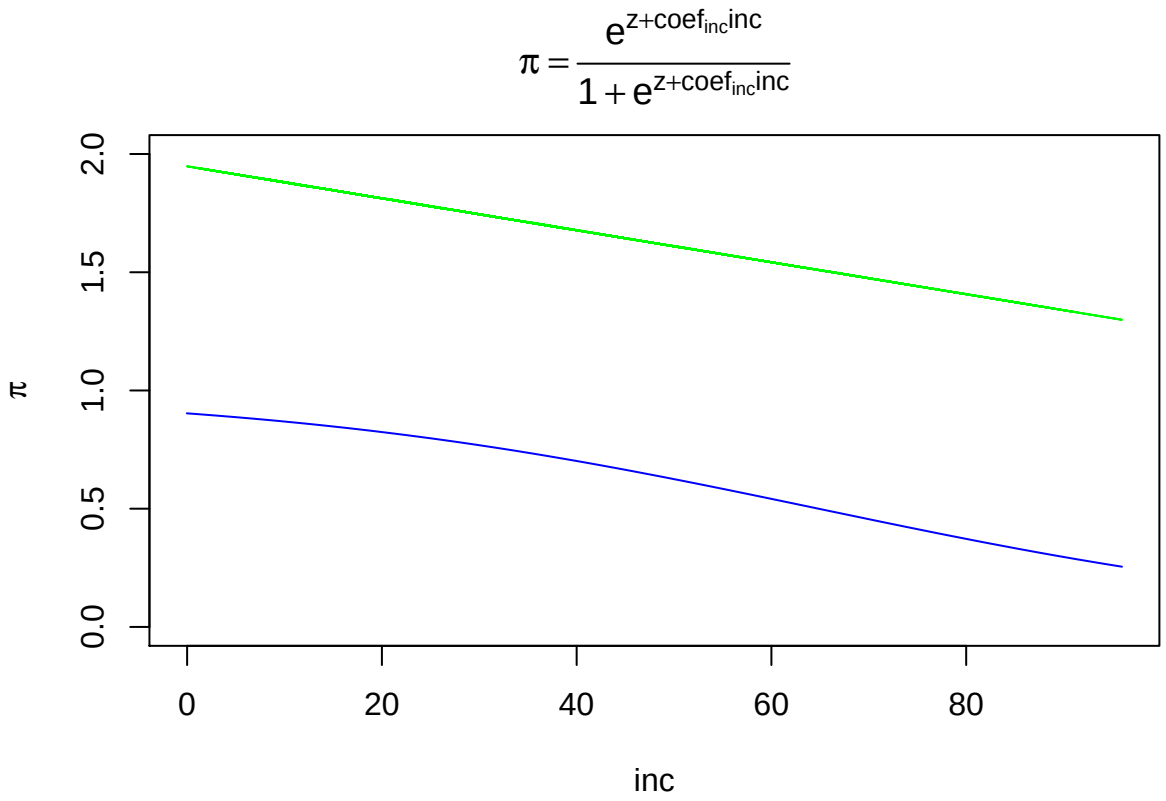
```
## -0.03444643
```

```
curve(expr = exp(z + coef[8] * x)/(1 + exp(z + coef[8] * x)),
    xlim = c(min(Mroz$inc), max(Mroz$inc)), ylim = c(0, 1), col = "blue",
    main = expression(pi == frac(e^{
        z + coef[inc] * inc
    }, 1 + e^{
        z + coef[inc] * inc
    })), xlab = expression(inc), ylab = expression(pi))
```

$$\pi = \frac{e^{z+coef_{inc}inc}}{1+e^{z+coef_{inc}inc}}$$




```
# Reproduce the graph overlaying the same result from the
# linear model as a comparison
curve(expr = exp(z + coef[8] * x)/(1 + exp(z + coef[8] * x)),
    xlim = c(min(Mroz$inc), max(Mroz$inc)), ylim = c(0, 2), col = "blue",
    main = expression(pi == frac(e^{
        z + coef[inc] * inc
    }, 1 + e^{
        z + coef[inc] * inc
    })), xlab = expression(inc), ylab = expression(pi))

par(new = TRUE)

y2 <- mroz.lm$coefficients[8] * x
lm.coef <- mroz.lm$coefficients
lm.z <- lm.coef[1] * xx[1] + lm.coef[2] * xx[2] + lm.coef[3] *
    xx[3] + lm.coef[3] * xx[3] + lm.coef[4] * xx[4] + lm.coef[5] *
    xx[5] + lm.coef[6] * xx[6] + lm.coef[7] * xx[7]

lines(x, lm.z + mroz.lm$coefficients[8] * x, col = "green")
```




```
# summary(mroz.lm) mroz.lm$coefficients[8]
```

Hypothesis testing and most importantly answering the questions

Testing Model Assumptions

1. The dependent variable needs to be binary (and not ordinal); specifically, the conditional distribution of y given follows a Bernoulli distribution
2. Observations are independent of each other. In fact, the error term of the model needs to follow an independent and identically distributed random variable.
3. No perfect collinearity
4. Linearity assumption: linearity of independent variables and log odds ratio

4. Other topics:

Model Parameter Estiamtion Algorithm: Iterated Reweighted Least Square

Reference: linked phrase