Live Session - Week 2: Discrete Response Models Lecture 2

Devesh Tiwari and Jeffrey Yau

September 12, 2017

Agenda

1. Q&A (estimated time: 5 minutes)
2. An overview of this lecture and live session (estimated time: 15 minutes)
3. An extended example (estimated time: 65 minutes)

1. Questions?

2. An Overivew of the Lecture

This lecture begins the study of logistic regression models, the most important special case of the generalized linear models (GLMs). It begins with a discussion of why classical linear regression models is not appropriate, from both statistical sense and practical application sense, to model categorical respone 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*

Recap some notations:

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

- the *link function* translates from the scale of mean response to the scale of linear predictor.

- The linear predicator can be expressed as

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

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

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

where $g()$ denotes some non-linear transformation. In the logit case, $g() = log_e(\frac{\mu}{1-\mu})$.

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

3. An extended example (estimated time: 65 minutes)

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)
```

Instructor's introduction to the example (estimated time: 5 minutes)

When solving data science problems, always begin with the understanding of the underlying question; our first step is typically **NOT** to jump right into the data. For the sake of this example, suppose the question is *"Do females who higher family income (excluding wife's income) have lower labor force participation rate?" If so, what is the magnitude of the effect?* 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.

Breakout Session 1: EDA. Time: 10 mins in groups. 5 mins discussion

Take a look at the dataset called *Mroz*, which is located in the *car* package in R. You can find a description of the variables in this dataset by typing ?Mroz in the R-editor. Answer the following questions about the EDA portion of the modelling process. Wherever possible, refer to the partial EDA included below as a guide; but more importantly, think about which questions an effective EDA should answer and how you would modify your modeling strategy based on those answers. Remember, the dependent variable here is dichotomous!

(1) What questions about the data are you trying to answer when you examine univariate plots? What are you looking for?

(2) What questions about the data are you trying to answer when you examine bivariate plots (between the dependent variable of interest and the independent variable and also between independent variables of interest)? What are you looking for?

(3) In many cases, an independent variable is continuous. How would you explore the relationship between this variable and a dichtomous DV? How would you be able to tell if you needed to include any non-linear transformation?

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

```
## Loading required package: dplyr

##
## 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
```

```r
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 ...
```

```r
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, ...
## $ 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...
```

```r
# 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
```

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

```
##     lfp k5 k618 age wc  hc       lwg    inc
## 93  yes  1    2  33 no yes 0.1148160 30.235
## 258 yes  1    0  53 no  no 1.2909843 18.275
## 278 yes  0    3  37 no  no 0.6208265 21.300
## 517  no  0    0  57 no yes 1.3051213 18.800
## 688  no  0    0  49 no  no 0.7900099 15.000
```

```r
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
```

```r
library(Hmisc)
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
## Loading required package: Formula
## Loading required package: ggplot2
##
## 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 distinct
##      753        0        2
##
## Value         no   yes
## Frequency    325   428
## Proportion 0.432 0.568
## ---------------------------------------------------------------------------
## k5
##        n  missing distinct     Info     Mean      Gmd
##      753        0        4    0.475   0.2377   0.3967
##
## Value          0     1     2     3
## Frequency    606   118    26     3
## Proportion 0.805 0.157 0.035 0.004
## ---------------------------------------------------------------------------
## k618
##        n  missing distinct     Info     Mean      Gmd
##      753        0        9    0.932    1.353     1.42
##
## Value          0     1     2     3     4     5     6     7     8
## Frequency    258   185   162   103    30    12     1     1     1
## Proportion 0.343 0.246 0.215 0.137 0.040 0.016 0.001 0.001 0.001
## ---------------------------------------------------------------------------
## age
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      753        0       31    0.999    42.54    9.289     30.6     32.0
##      .25      .50      .75      .90      .95
##     36.0     43.0     49.0     54.0     56.0
##
## lowest : 30 31 32 33 34, highest: 56 57 58 59 60
## ---------------------------------------------------------------------------
## wc
##        n  missing distinct
```

```
##      753        0        2
##
## Value         no   yes
## Frequency    541   212
## Proportion 0.718 0.282
## ---------------------------------------------------------------------------
## hc
##        n  missing distinct
##      753        0        2
##
## Value         no   yes
## Frequency    458   295
## Proportion 0.608 0.392
## ---------------------------------------------------------------------------
## lwg
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      753        0      676        1    1.097   0.6151   0.2166   0.4984
##      .25      .50      .75      .90      .95
##   0.8181   1.0684   1.3997   1.7600   2.0753
##
## lowest : -2.054124 -1.822531 -1.766441 -1.543298 -1.029619
## highest:  2.905078  3.064725  3.113515  3.155581  3.218876
## ---------------------------------------------------------------------------
## inc
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      753        0      621        1    20.13    11.55    7.048    9.026
##      .25      .50      .75      .90      .95
##   13.025   17.700   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
## ---------------------------------------------------------------------------
```

```r
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

Exercise (15 minutes): Instructor-led classwide discussion of the descriptive statistical analysis (or Exploratory Data Analysis)

An initiation of the descriptive statistical analysis:

- *Note that this descriptive statistics analysis is far from completed, and I leave it as take-home exercise for you to complete it. You are more than welcome to work with your classmates. Please volunteer to present your analysis next week.*

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. Also, we should not think of this list as exhaustive.

4. Summary of the discussion of univariate, bivariate, and multivarite analyses should come here. Note that most of these variables are categorical, making scatterplot matrix not an effective graphic device to visualize many bivariate relationships in one graph.

- Students to insert observations here. Discuss
  - the shape of the distribution, skewness, fat tail, multimodal, any lumpiness, etc
  - all of these distributional features across different groups of interest, such as number of kids in different age groups, husband's and wife's college attendance status
  - proportion of different categories
  - distribution in cross-tabulation (this is where contingency tables will come in handy)
- Think about engineering features (i.e. transformation of raw variables and/or creating new variables). Keep in mind that *log()* transformation is one of the many different forms of transformation. Note also that I use the terms *variables* and *features* interchangably. This lecture is a good place for you to review *w203*. For this specific dataset in this specific example, you may need to think about whether
  - to create a variable to describe the total number of kids?
  - to bin some of the variables? (Are some of the observations in some of the cell in the frequency or contingency tables too small?)
  - to creat spline function of some of the variables?
  - to transform one or more of the existing raw variables?
  - to create polynomial for one or more of the existing raw variables to capture non-linear effect?
  - to interact some of the variables?
  - to create sum or difference of variables?
  - etc

Take-home Exercises: Expand on the EDA I initiated below. Your analysis must be accompanied with detailed narrative.

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

```
## exp(Mroz$lwg)
##        n  missing distinct     Info     Mean      Gmd      .05      .10
```

```
##      753        0      676        1    3.567    2.236    1.242    1.646
##      .25      .50      .75      .90      .95
##    2.266    2.911    4.054    5.812    7.967
##
## lowest :  0.1282051  0.1616162  0.1709402  0.2136752  0.3571429
## highest: 18.2666721 21.4285726 22.5000020 23.4666673 25.0000019
```

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

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

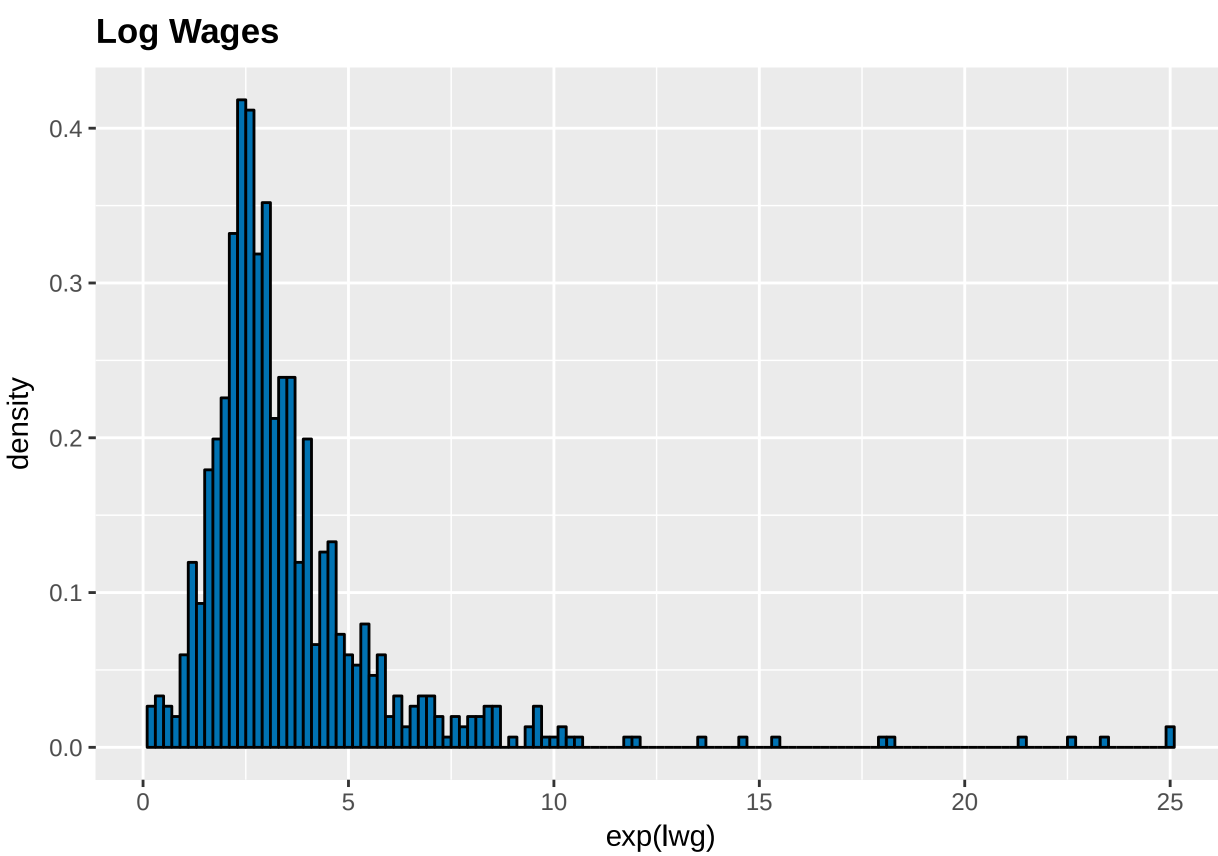
```r
require(ggplot2)
require(GGally)
```

```
## Loading required package: GGally
```

```
##
## Attaching package: 'GGally'
```

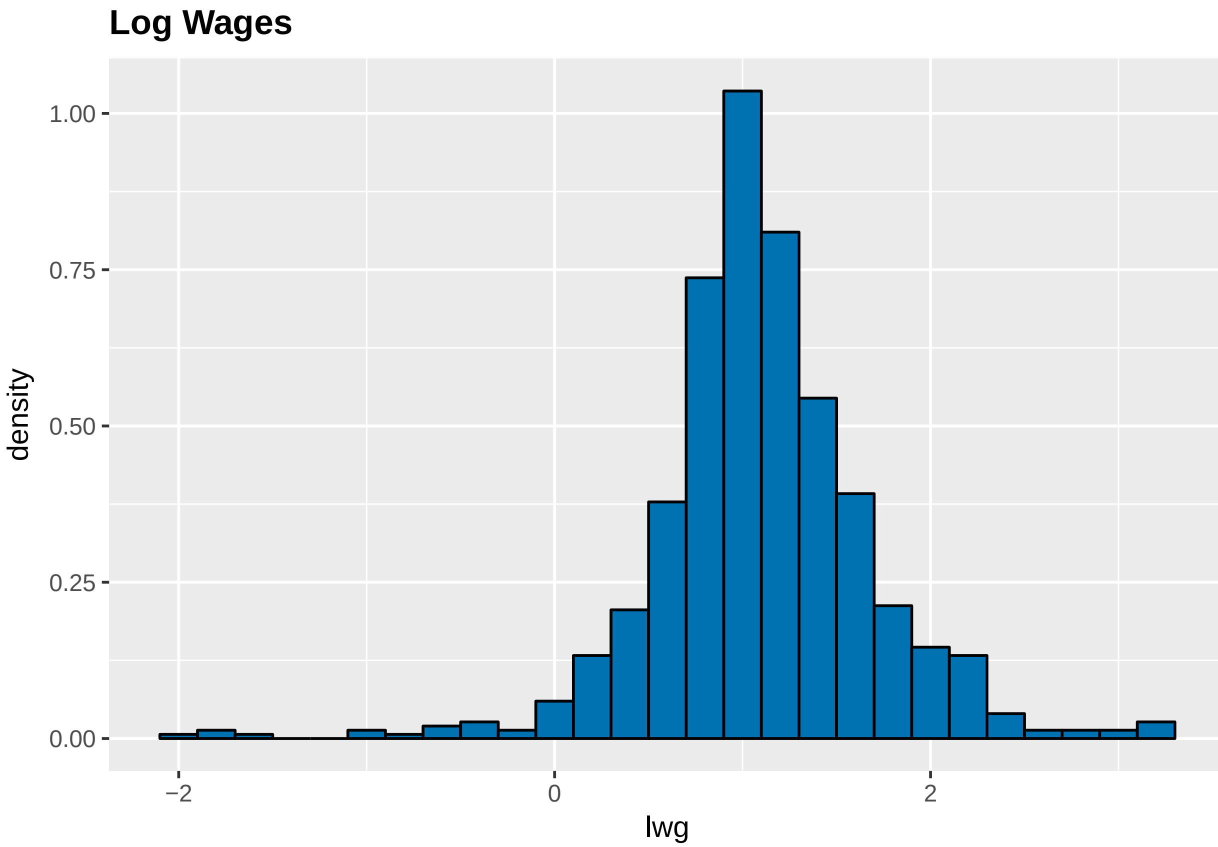
```
## The following object is masked from 'package:dplyr':
##
##     nasa
```

```r
## Distribution of wage Distribution of log(wage)
ggplot(Mroz, aes(x = exp(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"))
```

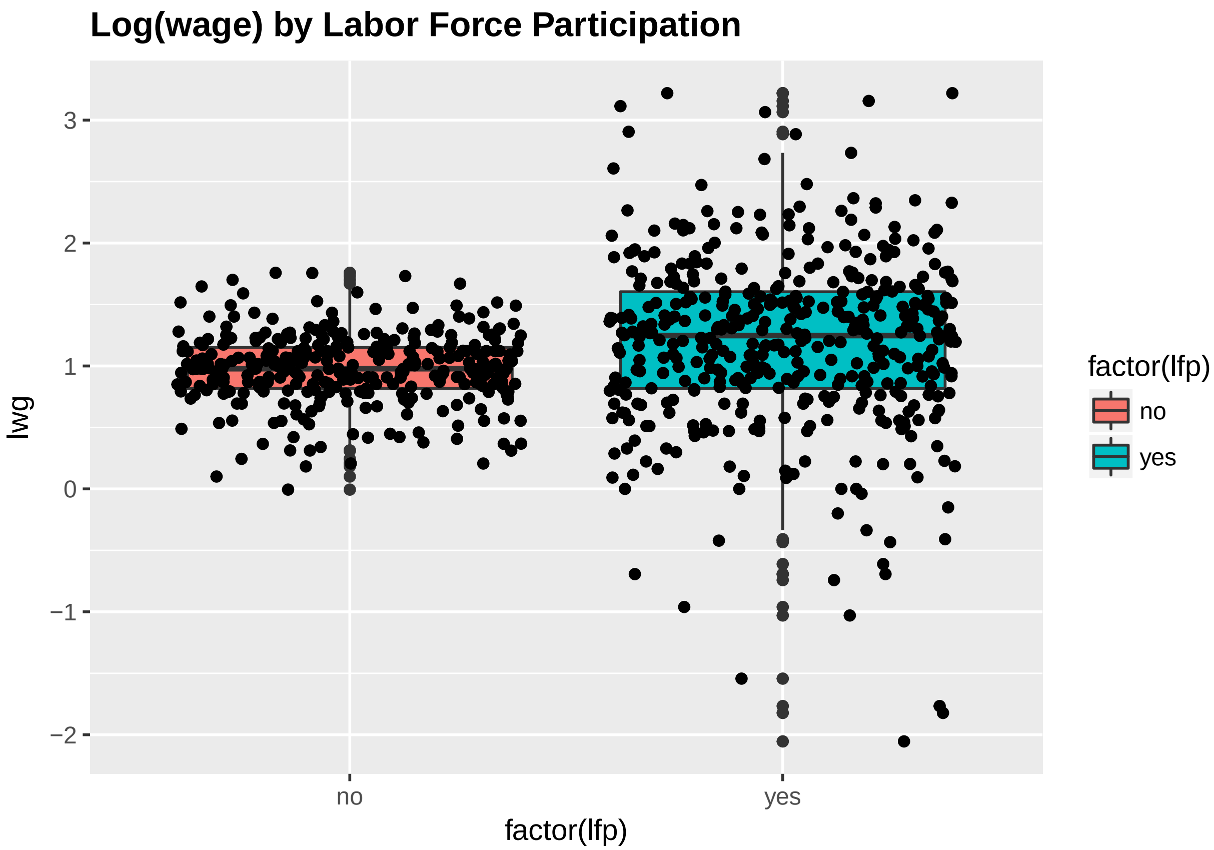



```r
# Distribution of log(wage)
ggplot(Mroz, aes(x = lwg)) + geom_histogram(aes(y = ..density..),
```

```
    binwidth = 0.2, fill = "#0072B2", colour = "black") + ggtitle("Log Wages") +
    theme(plot.title = element_text(lineheight = 1, face = "bold"))
```



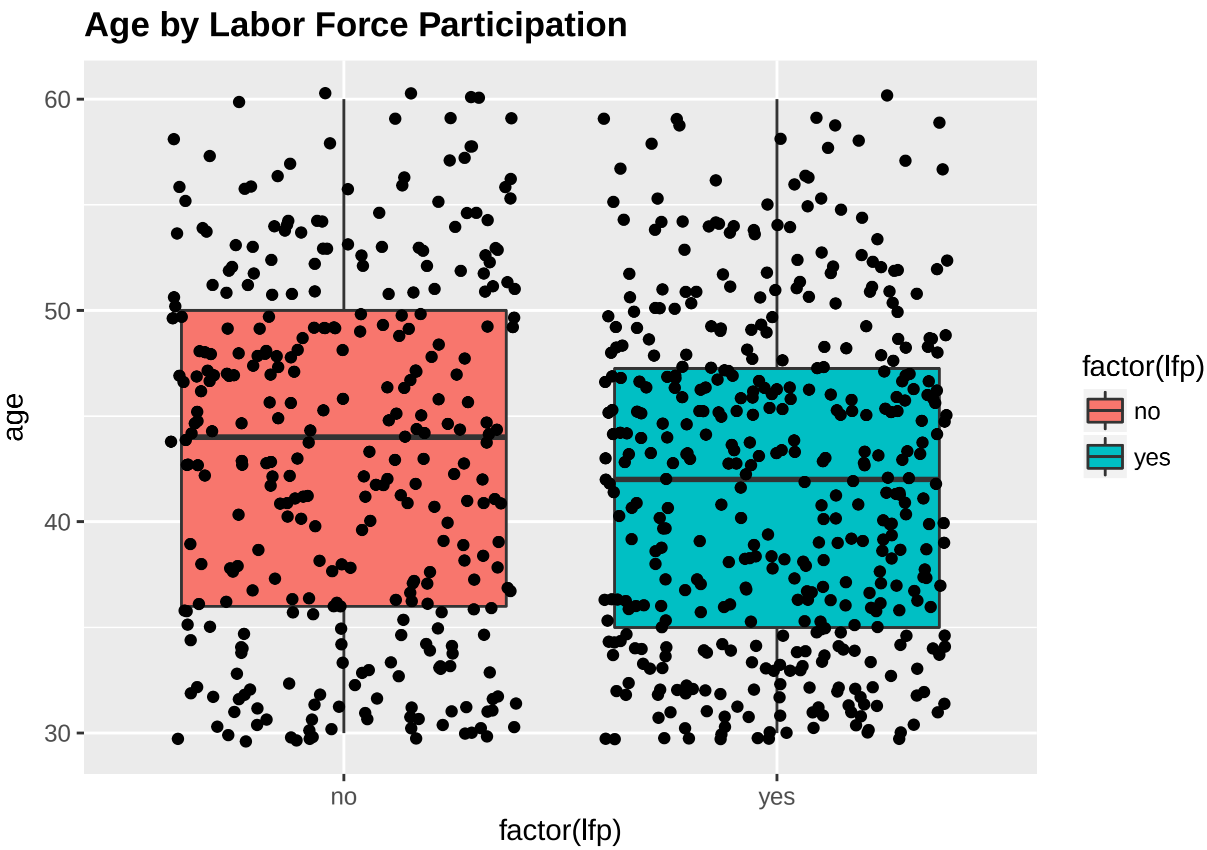

```
# log(wage) by lfp Do women who work earn more money?  Aside
# from mean value, what else is different?  Remember that if
# women do not work, their lwg is imputed.  What does this
# graph tell you?

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"))
```




```
t.test(Mroz$lwg ~ Mroz$lfp)
```

```
##
##  Welch Two Sample t-test
##
## data:  Mroz$lwg by Mroz$lfp
## t = -5.5951, df = 594.3, p-value = 3.369e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2912916 -0.1399265
## sample estimates:
##  mean in group no mean in group yes
##         0.9745642         1.1901732
```

```
# 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"))
```



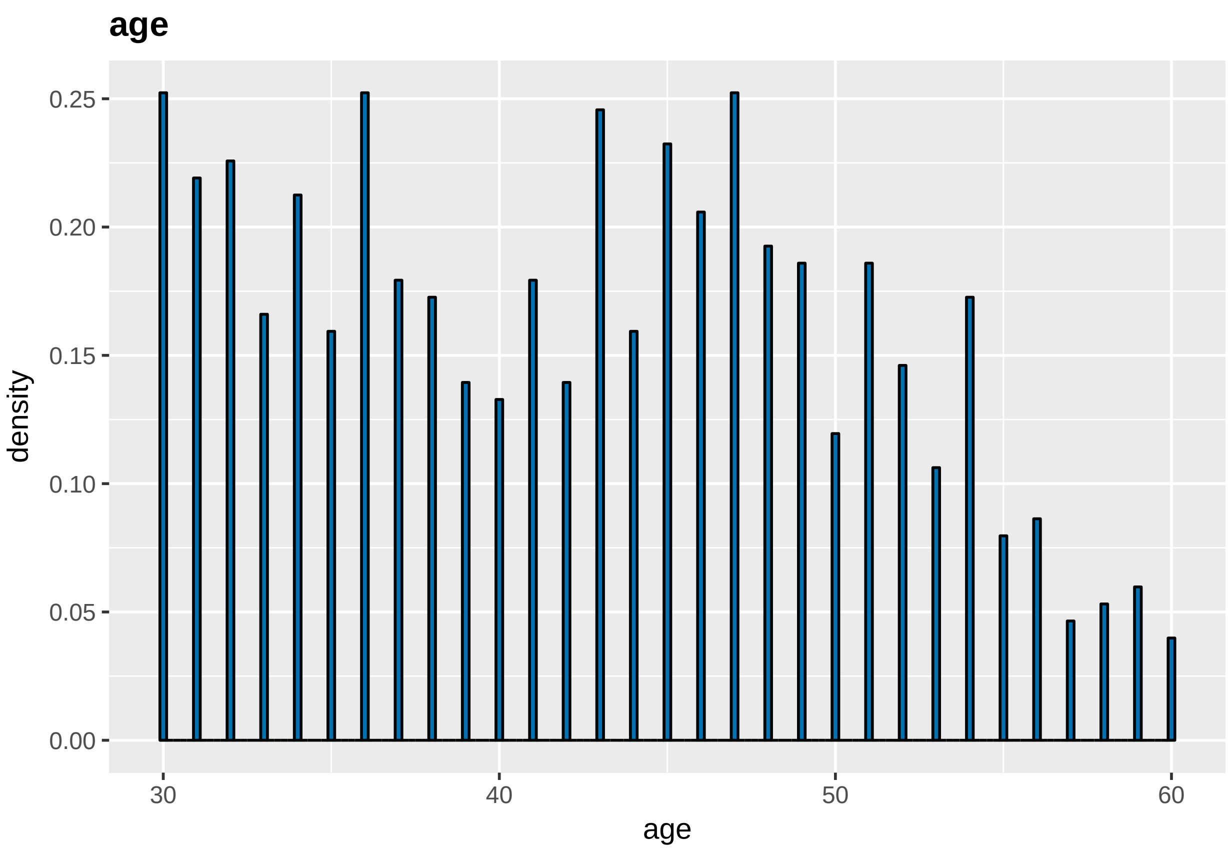

```r
t.test(Mroz$age ~ Mroz$lfp)
```

```
##
##  Welch Two Sample t-test
##
## data:  Mroz$age by Mroz$lfp
## t = 2.1855, df = 662.02, p-value = 0.02921
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1331226 2.4891060
## sample estimates:
##  mean in group no mean in group yes
##          43.28308          41.97196
```

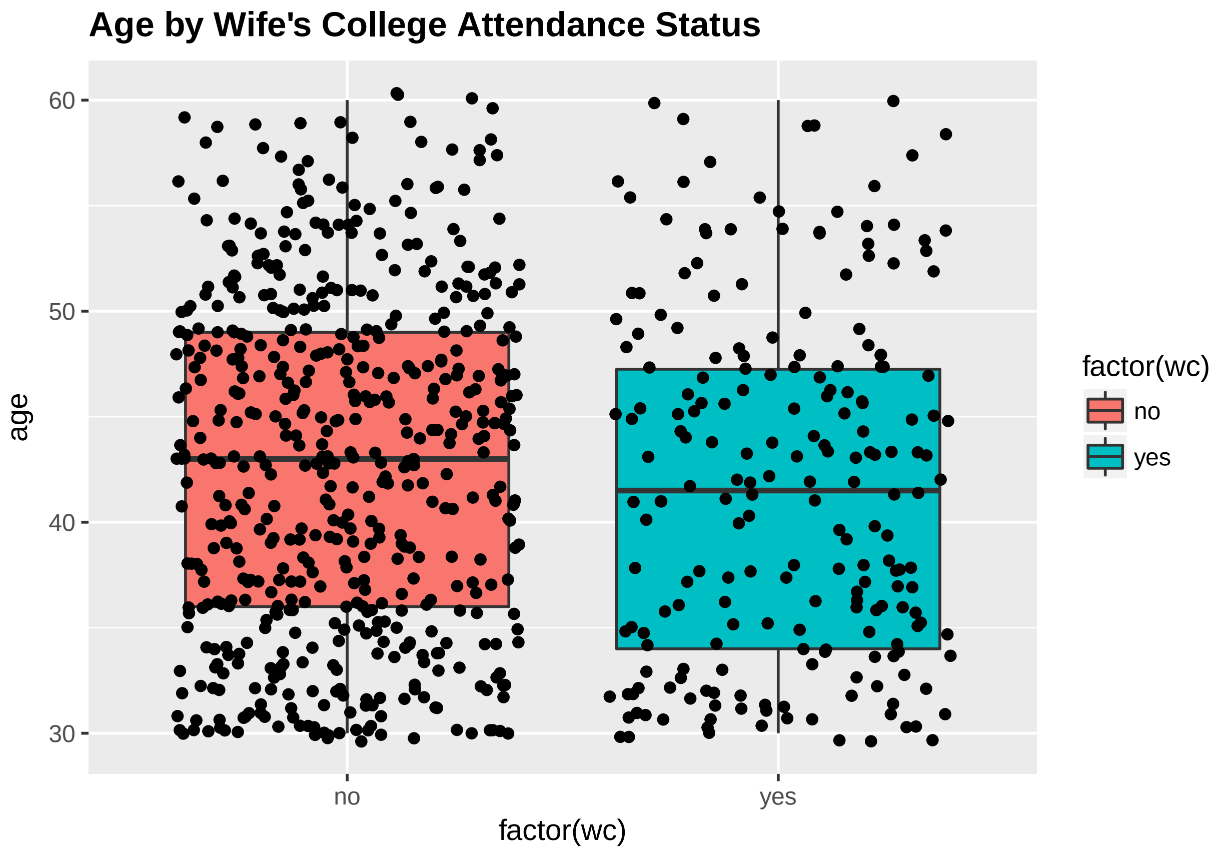
```r
# Distribution of age
summary(Mroz$age)
```

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

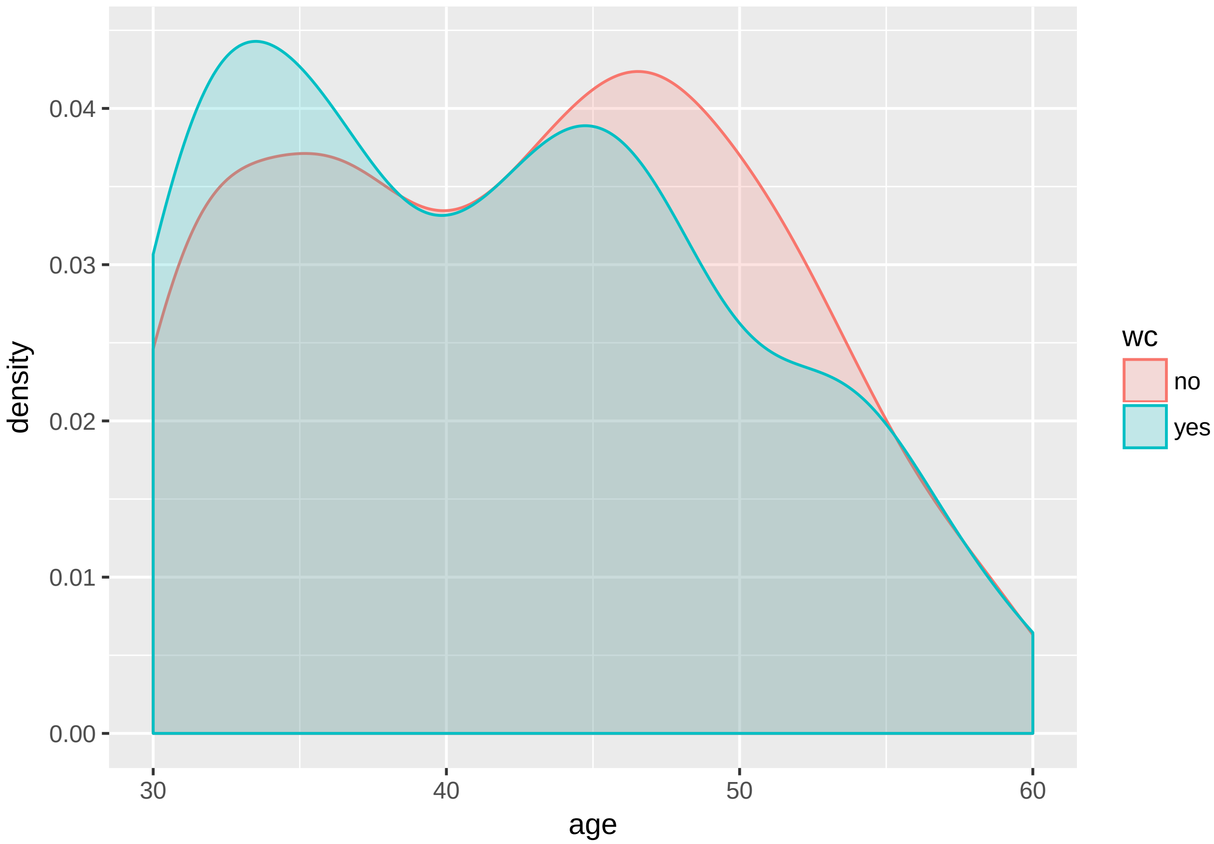
```r
ggplot(Mroz, aes(x = age)) + geom_histogram(aes(y = ..density..),
    binwidth = 0.2, fill = "#0072B2", colour = "black") + ggtitle("age") +
    theme(plot.title = element_text(lineheight = 1, face = "bold"))
```



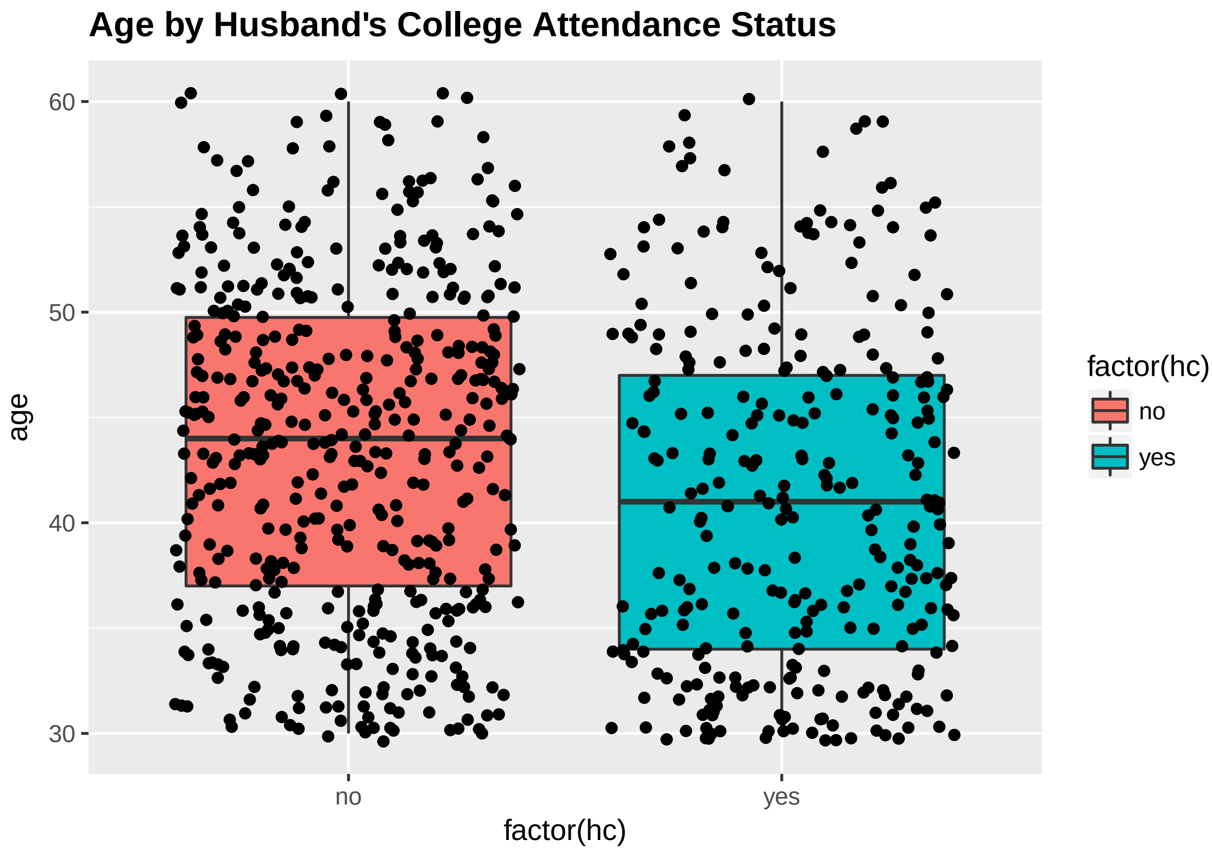

```
## Any observations here?


# Distribution of age by wc Were those who attended colleage
# tend to be younger?  If so, what does that tell us?
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"))
```



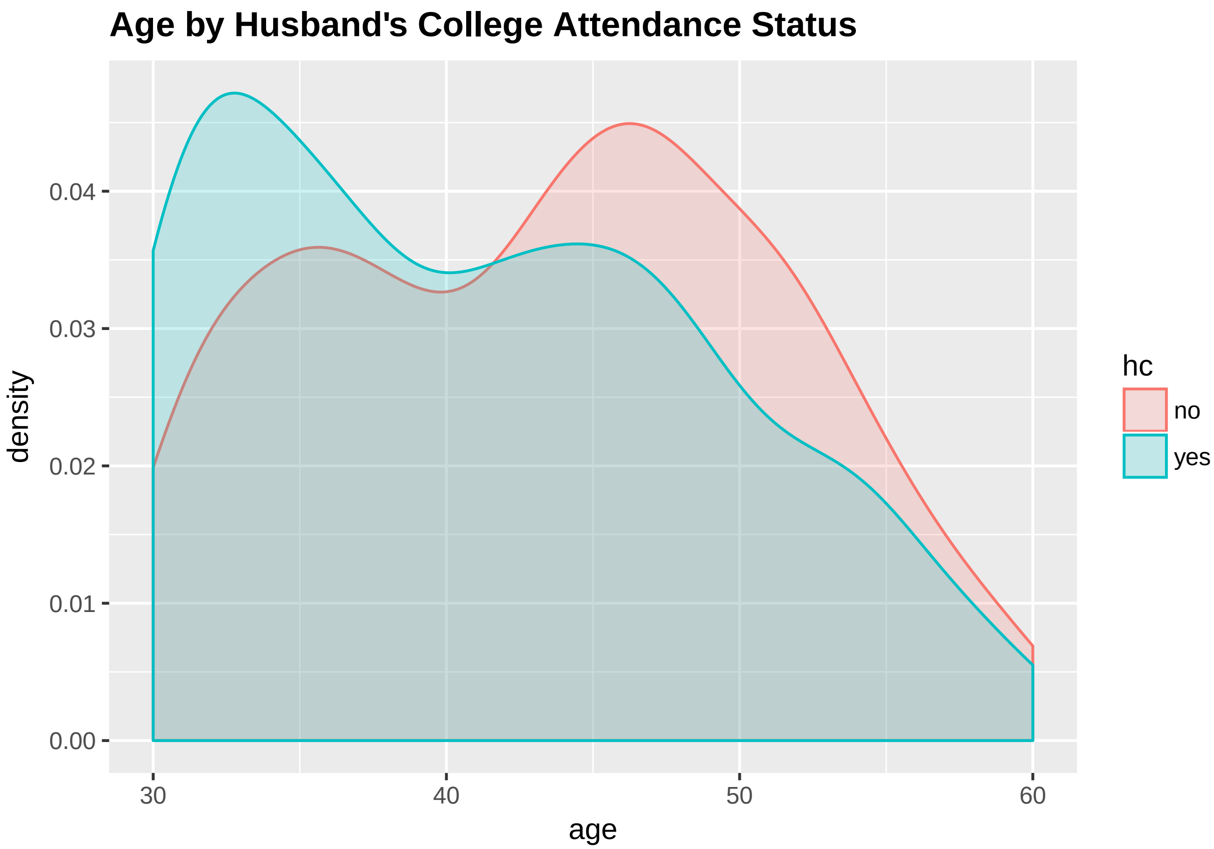

```r
# Sometimes it is usefyl to examine the underlying
# distribution of a variable in each category
ggplot(Mroz, aes(age, fill = wc, colour = wc)) + geom_density(alpha = 0.2)
```



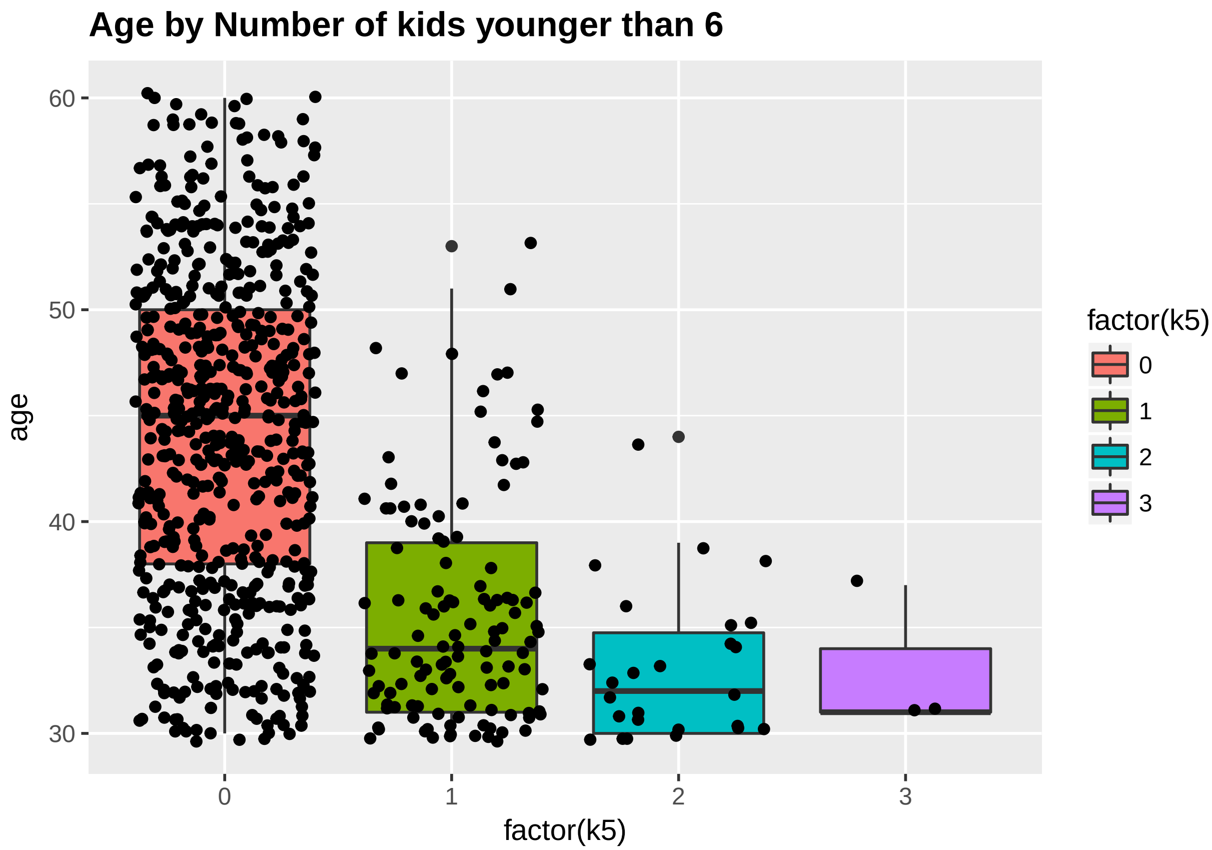

```r
# Distribution of age by hc Were those whose husband attended
# colleage tend to be younger?
ggplot(Mroz, aes(factor(hc), age)) + geom_boxplot(aes(fill = factor(hc))) +
    geom_jitter() + ggtitle("Age by Husband's College Attendance Status") +
    theme(plot.title = element_text(lineheight = 1,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



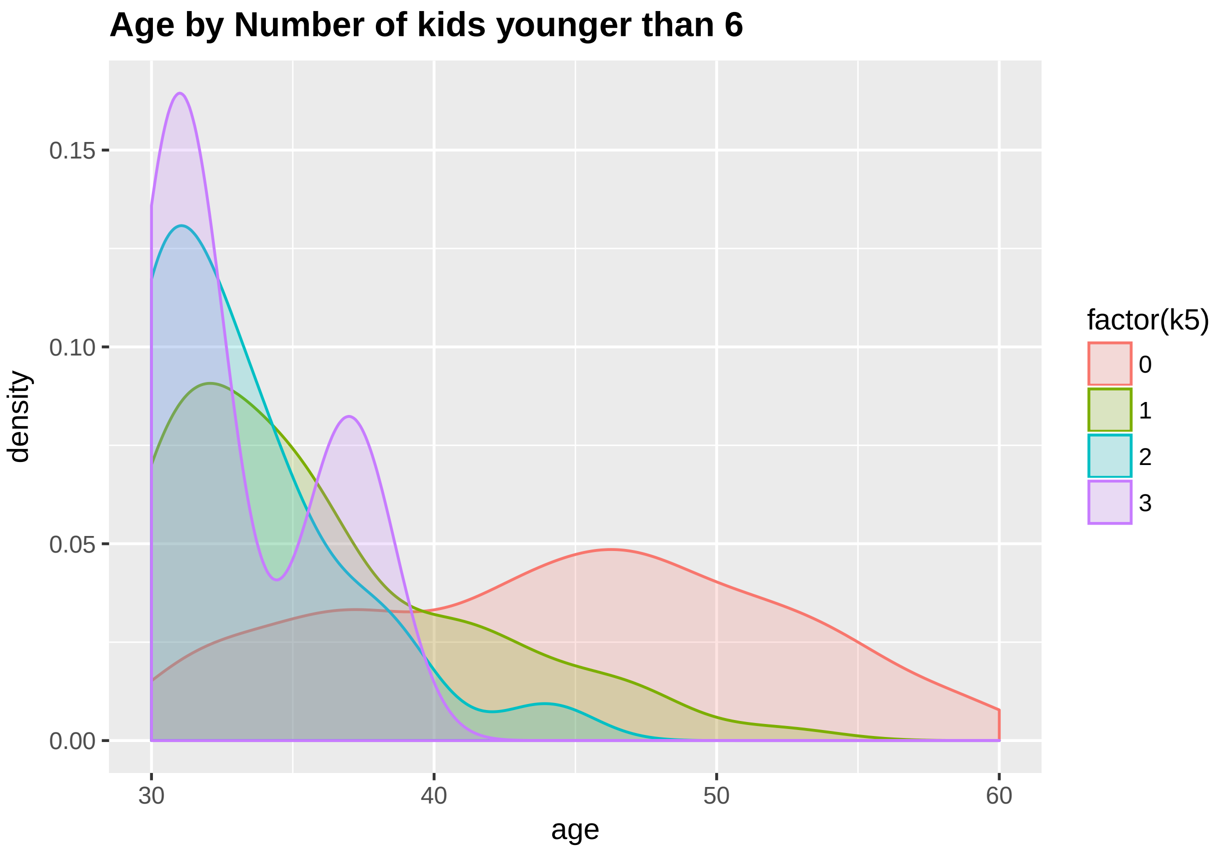

```
ggplot(Mroz, aes(age, fill = hc, colour = hc)) + geom_density(alpha = 0.2) +
    ggtitle("Age by Husband's College Attendance Status") + theme(plot.title = element_text(lineheight
    face = "bold"))
```



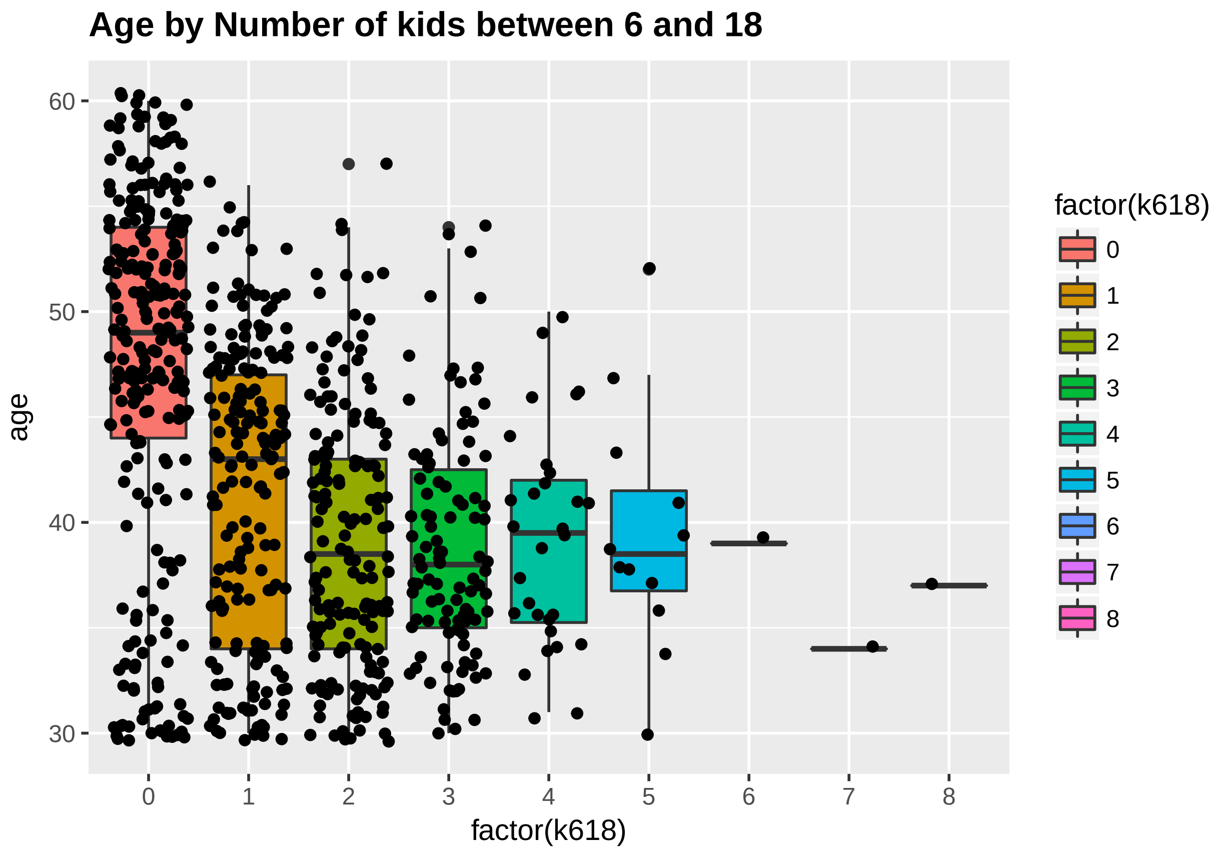

```r
# Distribution of age by number kids in different age group
ggplot(Mroz, aes(factor(k5), age)) + geom_boxplot(aes(fill = factor(k5))) +
    geom_jitter() + ggtitle("Age by Number of kids younger than 6") +
    theme(plot.title = element_text(lineheight = 1, face = "bold"))
```



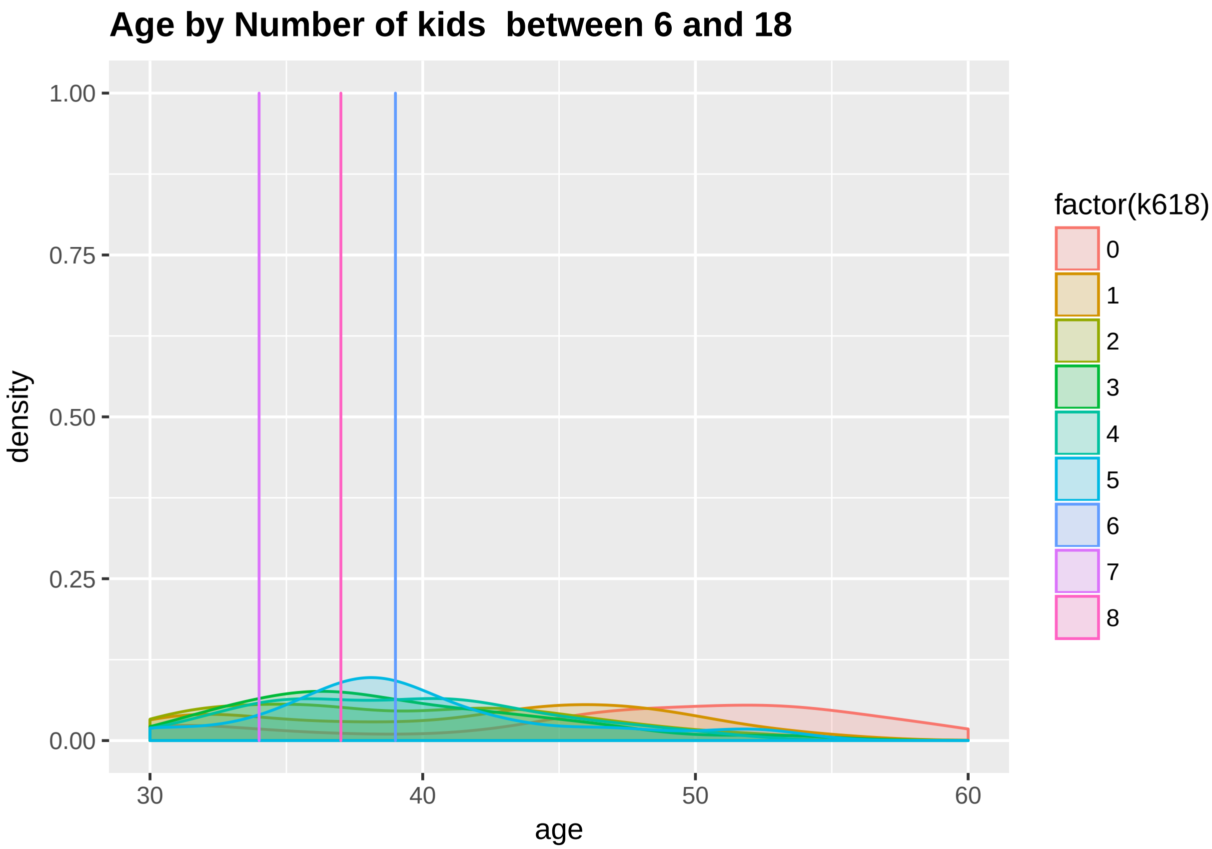

```r
# Is this surprising?

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"))
```




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

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

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

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

```r
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
```

```r
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
```

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

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

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

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

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

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

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

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

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

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

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

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

```r
# Anything intersting here?
```

As a best practice, we will need to incorporate insights generated from EDA on model specification. As you see below, I will assign it as take-home exercise. In what follows, I employ a very simple specification that uses all the variables as-is.

Group Discussion: Comparing a linear model with a logit model.

In this exercise, we are going to examine the relationship between the dependent variable, *lfp*, and the remaining covariates via the CLM and logistic regression. Please follow the steps below as described:

(1) I built a linear model in the code below and a logistic regression. Interpret the impact of the variable *k5* on *lpv* for both models. Pay attention to the distribution of *k5*, what it stands for, and what the coefficient itself tells us. Think about whether or not you would code *k5* any differently.

```r
library(stargazer)
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2015). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2. http://CRAN.R-project.org/package=stargazer
```

```r
mroz.lm <- lm(as.numeric(lfp) - 1 ~ k5 + k618 + age + wc + hc +
    lwg + inc, data = Mroz)


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


stargazer(mroz.lm, mroz.glm, type = "text")
```

```
##
## =====================================================
##                            Dependent variable:
##                      --------------------------------
##                        as.numeric(lfp) - 1     lfp
##                               OLS            logistic
##                               (1)              (2)
## -----------------------------------------------------
## k5                         -0.295***        -1.463***
##                             (0.036)          (0.197)
##
## k618                        -0.011           -0.065
##                             (0.014)          (0.068)
##
## age                        -0.013***        -0.063***
##                             (0.003)          (0.013)
##
## wcyes                      0.164***         0.807***
##                             (0.046)          (0.230)
##
## hcyes                        0.019            0.112
##                             (0.043)          (0.206)
##
## lwg                        0.123***         0.605***
##                             (0.030)          (0.151)
##
## inc                        -0.007***        -0.034***
##                             (0.002)          (0.008)
##
## Constant                   1.144***         3.182***
##                             (0.127)          (0.644)
##
## -----------------------------------------------------
## Observations                  753              753
## R2                           0.150
## Adjusted R2                  0.142
## Log Likelihood                              -452.633
## Akaike Inf. Crit.                            921.266
## Residual Std. Error    0.459 (df = 745)
## F Statistic         18.827*** (df = 7; 745)
## =====================================================
```

```
## Note:                           *p<0.1; **p<0.05; ***p<0.01
```

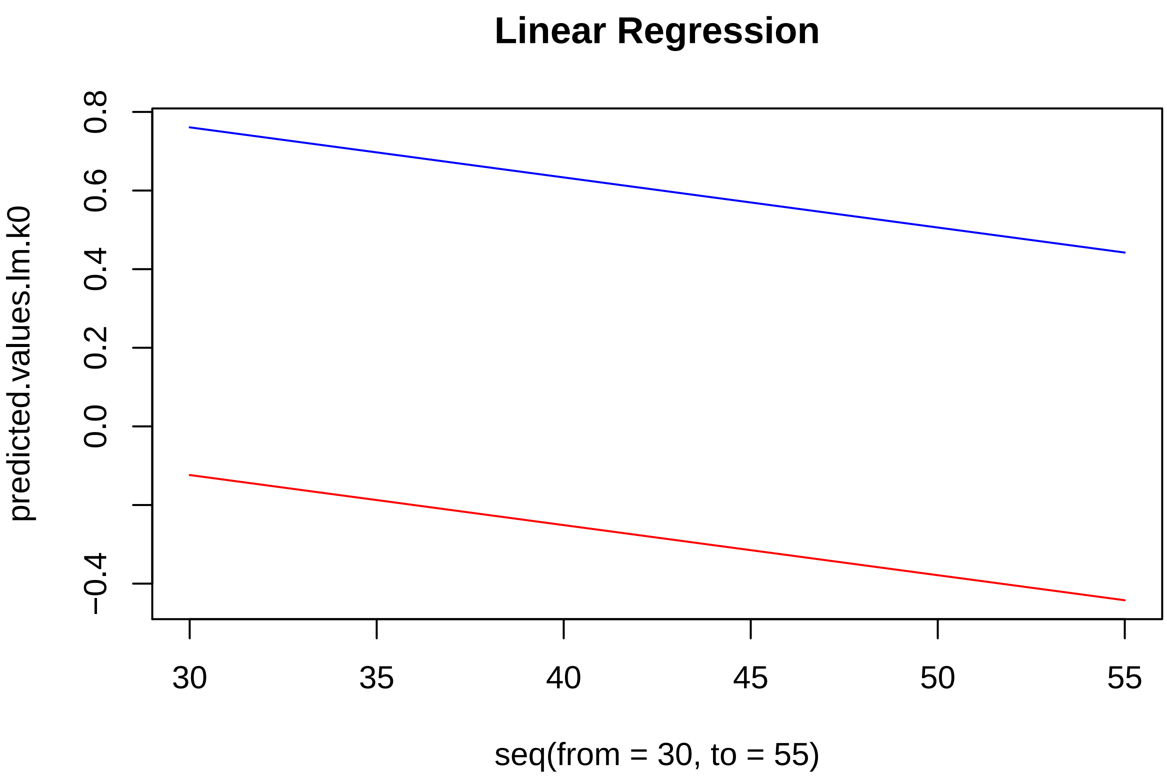
(2) Let's visually examine the relationsip between age and lfp for both the CLM and logistic models across two scenarios: One where *k5* equals zero and another when it equals three. In order to do this, we will need to use the predict.lm and the predict.glm functions in R. Take a minute to look at the documentations, but these two functions use our model results to generate predicted values on values specified by the user (see my code below on how to do that).

All told, you will generate 4 sets of predicted values, two for the clm model and two for the logit model. Plot all four of these predicted values against age (you don't have to do it all in a single plot, for now do what is easiset for you).

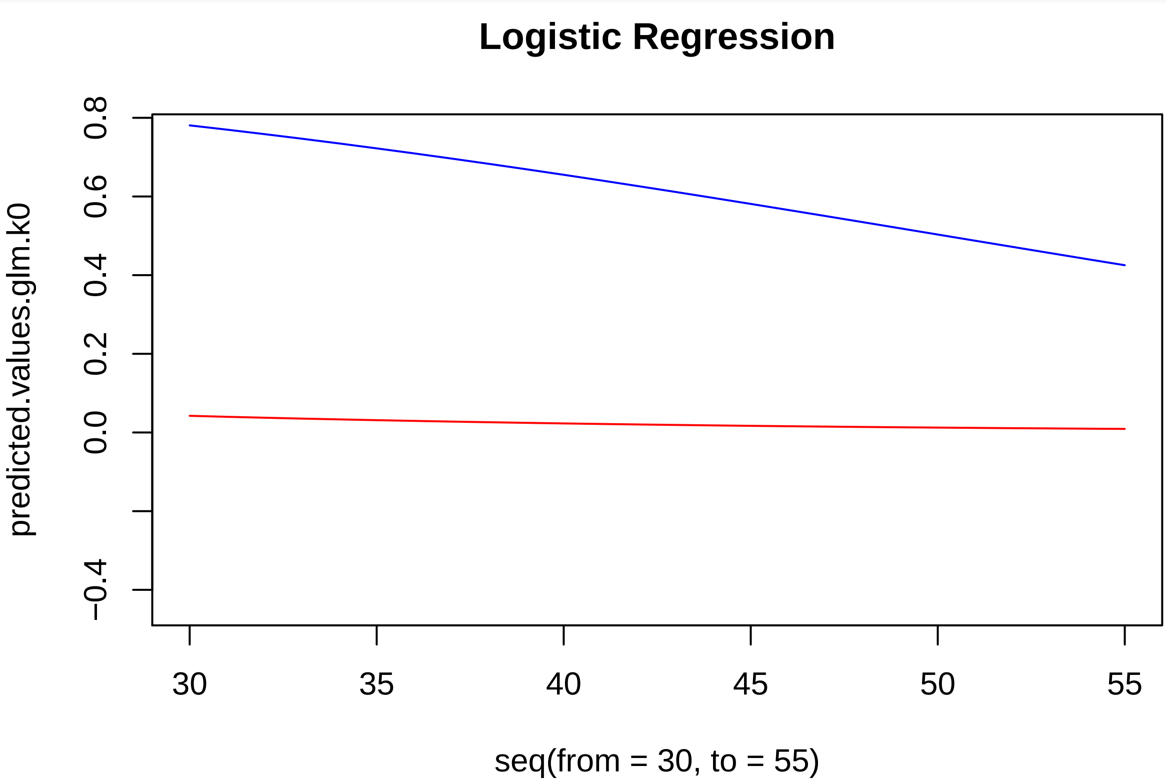
For this exercise, do not worry about the confidence intervals — we will tackle those next week.

Examine the plots and note anything that looks interesting or note-worthy. We will talk about this togther.

```r
# Create the new df that will be used by the predict
# functions.  You will use this df for both the predict.lm
# and predict.glm functions

newdf <- data.frame(k5 = 0, k618 = 0, age = seq(from = 30, to = 55),
    wc = "no", hc = "no", lwg = 1.0971, inc = 20)

predicted.values.lm.k0 <- predict.lm(mroz.lm, newdata = newdf,
    se.fit = FALSE)
predicted.values.glm.k0 <- predict.glm(mroz.glm, newdata = newdf,
    type = "response")


newdf <- data.frame(k5 = 3, k618 = 0, age = seq(from = 30, to = 55),
    wc = "no", hc = "no", lwg = 1.0971, inc = 20)

predicted.values.lm.k3 <- predict.lm(mroz.lm, newdata = newdf,
    se.fit = FALSE)
predicted.values.glm.k3 <- predict.glm(mroz.glm, newdata = newdf,
    type = "response")


# Plots. LM
plot(x = seq(from = 30, to = 55), predicted.values.lm.k0, type = "l",
    col = "blue", ylim = range(c(predicted.values.lm.k0, predicted.values.lm.k3)),
    main = "Linear Regression")
lines(x = seq(from = 30, to = 55), predicted.values.lm.k3, col = "red")
```




```r
# Plots Logistic regression
plot(x = seq(from = 30, to = 55), predicted.values.glm.k0, type = "l",
     col = "blue", ylim = range(c(predicted.values.lm.k0, predicted.values.lm.k3)),
     main = "Logistic Regression")
lines(x = seq(from = 30, to = 55), predicted.values.glm.k3, col = "red")
```

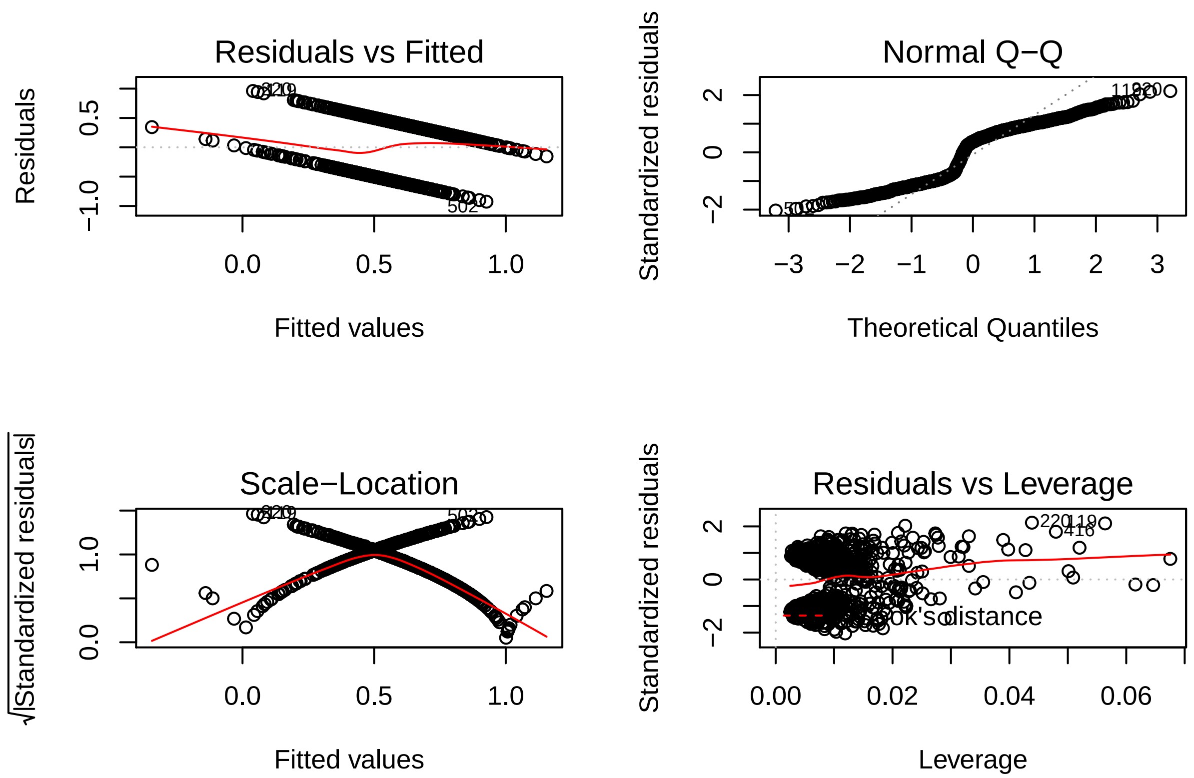


```
#### Question: If you did not have the predict function, how
#### would you have constructed the plots for the logistic
#### function?
```

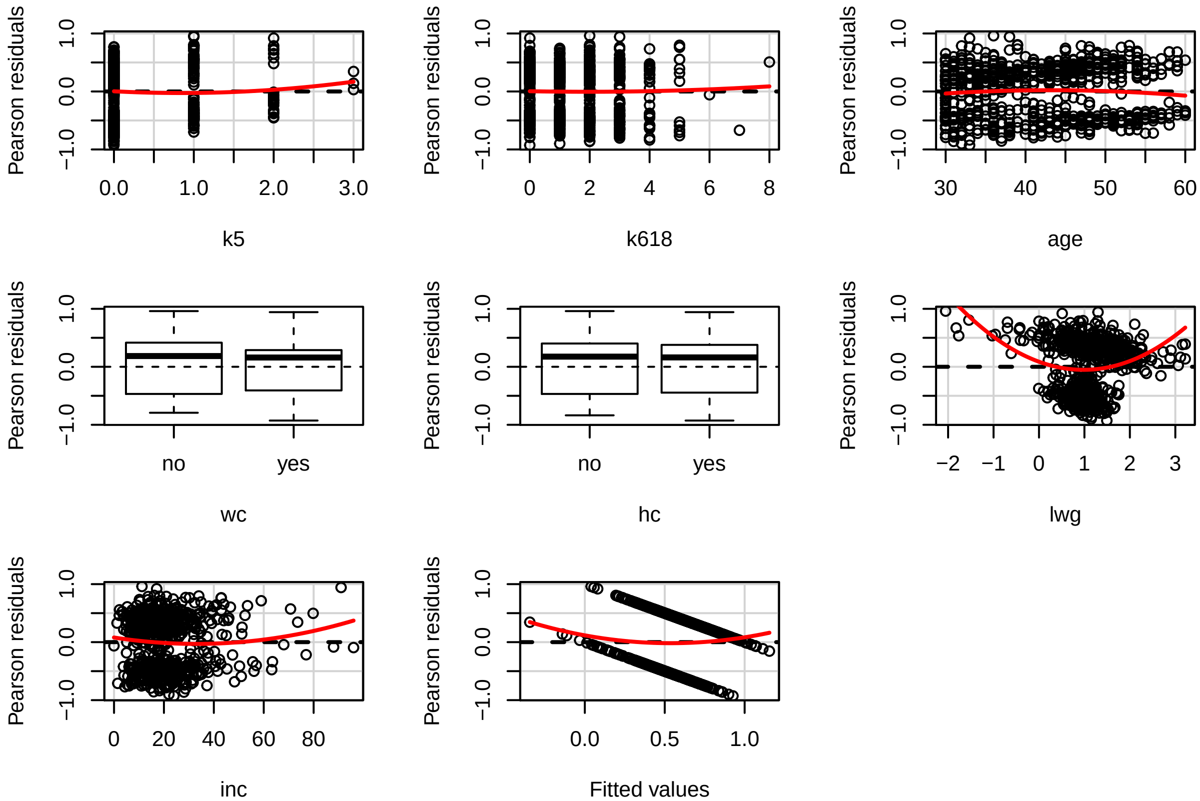
Exercise: Residual analysis

THIS IS FOR ILLUSTRATION PURPOSES ONLY!! Suppose we conducted the same type of residual analysis as we would have under the CLM. Review the plots below. What do you notice? Are there any shortcomings with using this type of residual analysis?

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




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



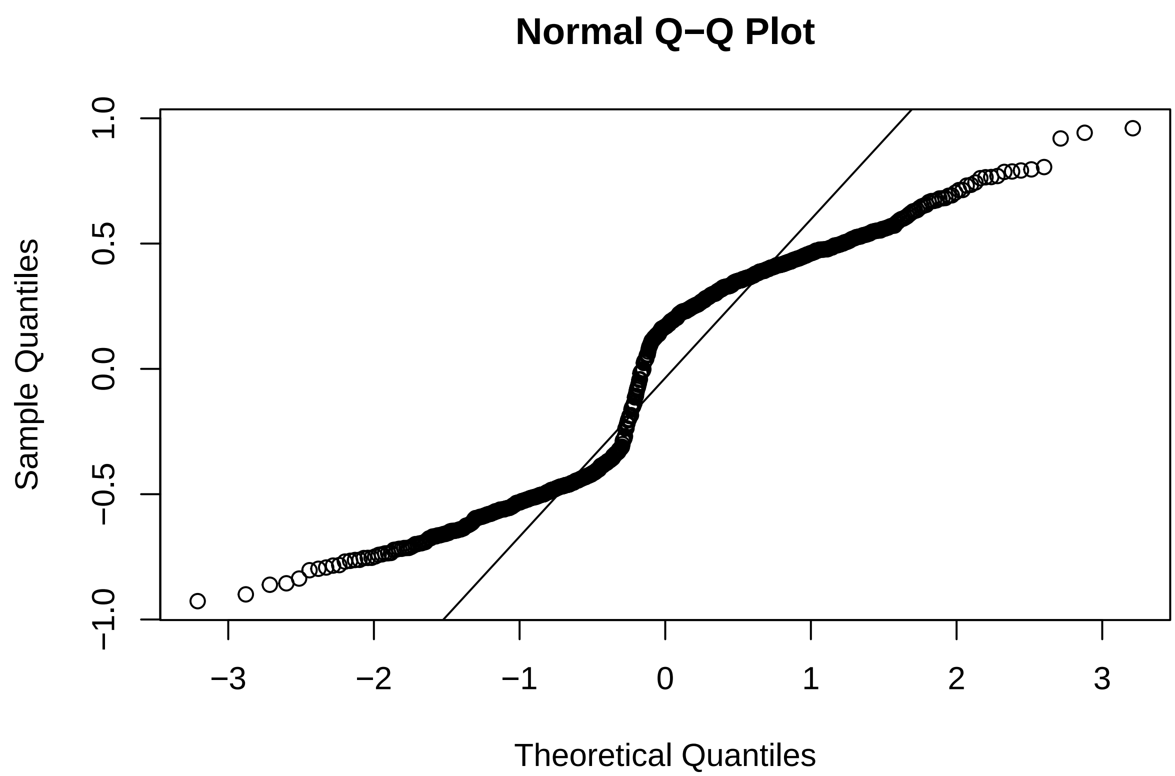

```
##              Test stat Pr(>|t|)
## k5               0.969    0.333
## k618             0.384    0.701
## age             -1.347    0.178
## wc                  NA       NA
## hc                  NA       NA
## lwg              7.697    0.000
## inc              1.970    0.049
## Tukey test   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

```r
# Note that I didn't pay much attention to outliers and
# influential observations in this specific example, but you
# should comment on it.

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.3442  0.4555  0.5681  0.5684  0.6987  1.1550
```

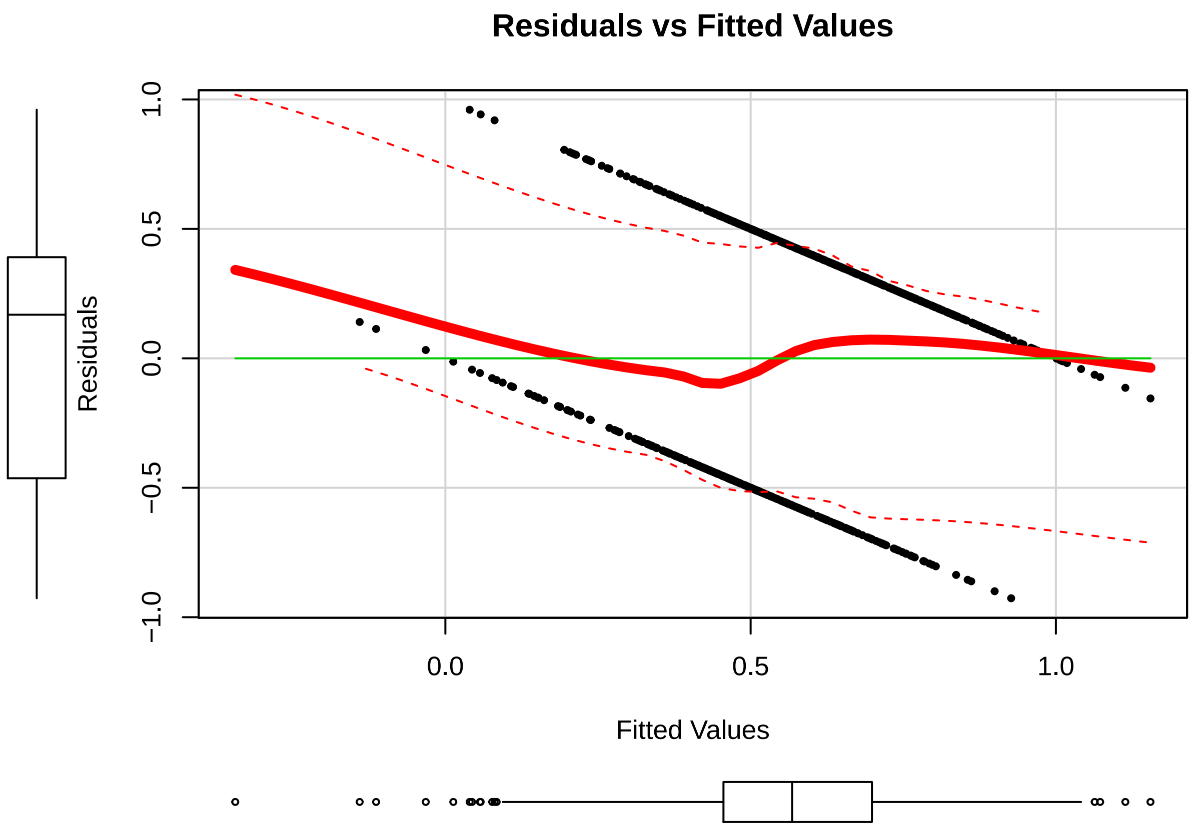
```r
# par(mfrow=c(1,1)) plot(mroz.lm$residuals,
# main='Autocorrelation Function of Model Residuals')
# acf(mroz.lm$residuals, main='Autocorrelation Function of
# Model Residuals')

hist(mroz.lm$residuals)
```


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




```r
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")
```



Take-home exercises

1. Use the model *mroz.glm* and test the hypothesis the hypothesis the wife's wage had no impact on her labor force participation. Set up the test. Write down the null hypothesis. Explain which test(s) you used. State the results. Explain the results.
2. Explain all of the deviance statistics in the model results (*summary(mroz.glm)*) and what do they tell us? (You answer may require you to perform further calculation using the deviance statistics.)
3. Expand the EDA and propose one additional specification based on your EDA.
4. Test this newly proposed model, call it mroz.glm2, and test the difference between the two models.
5. Study the model parameter estiamtion algorithm: Iterated Reweighted Least Square (IRLS) Reference: linked phrase