Live Session - Week 2: Discrete Response Models Lecture 2

Dr. Jeffrey Yau

5/11/2021

Introduction

Agenda

	Estimated Time	Topics
1	5 minutes	Folks coming in
2	10 minutes	Lecture Overview
3	10 minutes	A Quick Review of Linear Probability Model
4	10 minutes	A Quick Review of Binary Logistic Regression Model
5	45 minutes	A walk-through of a guided example

An Overivew of the Lecture

Required Readings: BL2015: Ch. 2.1, 2.2.1 - 2.2.6

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 (its advantages, and its limitations), 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

Learning Objectives

In this lecture, students will learn

- The mathematical formulation of Binary Response Models, Linear Probability Model, its advantages, and its limitations
- Common non-linear transformation used in the context of binary dependent variable
- Binary Logistic Regression Model
- Underlying assumptions of Binary Logistic Regression Model
- Maximum likelihood estimation and an overview of a numerical procedure used in practice
- Variance-Covariance matrix of the estimates
- Hypothesis testing
- Discusses how to estimate and make inferences about a single probability of success
- The notion of deviance
- Odds ratios in the context of binary logistic regression model
- Discussion of probability of success and its associated inference
- Visual assessment of logistic regression model

Regression Models of Binary Response Variable

Bernoulli and Binomial Probability Models

Recall from *w203* the *Bernoulli and Binomial Probability Models*, in which these models are not "tied" to any explanatory variables used to model the "relationship" between the probability (or some functions of the probability) with these variables, and the parameters of those probability "model" can be "estimated" using a sample of data.

Consider a trial whose outcome can be classified as either a success or failure (or some event occurs or does not occur). Define a random variable X that takes the value of 1 if the trial is a success and 0 if it is a failure. Then, the probability mass function of X is given by

$$p(0) = P(X = 0) = 1 - p$$
$$p(1) = P(X = 1) = p$$

where p, $0 \leq p \leq 1$, denotes the probability that the trial is a success. In this case, the random variable X is called a Bernoulli random variable (after the Swiss mathematician James Bernoulli).

Extending this idea, let's say we have n independent trials, each of which results in a success with probability p and in a failure with probability $1 - p$.

What do you notice in this statement regarding an implicit assumption regarding the "distribution" followed by these trials?

Now, let's use X to represent the number of successes in n trials. Then, X is said to be a *Binomial random variable* with parameters (n, p). (Side note: if one is not careful about terminology used in different context, it is a very dangerous situation when brought into the machine learning domain, as it may lead people to consider p as a parameter!) Notice that *Bernoulli* random variable is a special case of *Binomial* random variable with $n = 1$.

The probability mass function of a binomial random variable with parameters (n, p) is given by

$$p(i) = \binom{n}{i} p^i (1-p)^{(n-i)} \qquad i = 0, 1, \ldots, n$$

* **How do we estimate p?**

- **How do we estimate p as a function of a set of explanatory variables?**

Linear Probability Model

Given a set of n realizations from K explanatory variables, $\{x_{i1}, \ldots x_{iK}\}$, a regression model relates the dependent variable, $P(Y = 1) = \pi$, with the set of explanatory variables via a parametric function $g()$ with the parameters β:

$$\pi_i = P(Y_i = 1 | x_{i1}, \ldots x_{iK}) = g(x_{i1}, \ldots x_{iK} | \beta)$$

Different functional forms of $g()$ give different regression models.

If $g()$ is an linear function, then we have a *linear probability model*, which has many drawbacks and should not be used:

$$\pi_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_K x_{iK} + \epsilon_i$$

Breakout room discussion:

- **What are the advantages of the linear probability model?**

- **What are the drawbacks of the linear probability model?**
- **Have you you the linear probability model in your work or in other context? If so, please describe the situation in which the linear probability model is applied.**

Binary Logistic Regression

Formulation

$$\begin{aligned}\pi_i &= P(Y_i = 1|x_{i1}, \dots x_{iK}) \\ &= g(x_{i1}, \dots x_{iK}|\beta) \\ &= \frac{exp(z_i)}{1 + exp(z_i)}\end{aligned}$$

where

$$z_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_K x_{iK}$$

- 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

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

Another way to express a logistic regression is

$$logit\,(\pi_i) = log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_K x_{iK}$$

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)
```

Practical Tips for Implementing Binary Logistic Regression

When solving data science problems, always begin with the understanding of the underlying (business, policy, scientific, etc) question; our first step is typically **NOT** to jump right into the data.

For this example, suppose the question is *"Do females who have 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 objective in *Mroz (1987)'s* 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 data: Remember that this sample comes from *1976 Panel Data of Income Dynamics (PSID)*. PSID is one of the most popular datasets used by labor economists.

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

Typical questions you should always ask when examining a dataset 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 these variables sufficient for you to answer you questions?
- If not, what other variables would you like to have? What impact (qualitatively) might not having these variables have on your models?
- What are the number of observations?
- Are there any missing values (in each of the variables)?
- Are there any abnormal values in each of the variables in the raw data?

Note: in practice, you will likely query your data from different tables potentially from different databases, clearn them, process them, join them, and perhaps process them even further. This is before any feature engineering step. However, we will not do any of these in this course.

```
# Import libraries
library(car)
```

```
## Loading required package: carData
```

```
library(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
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':
##
##     src, summarize

## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```r
# Set working directory
# setwd("~/Documents/Teach/Cal/w271/course-main-dev/live-session-files/week02")
wd <- getwd()
wd
```

```
## [1] "/Users/FK/Documents/Work/Cal/w271/main-2021-fall/live_session/live_session_02"
```

```r
?Mroz
data(Mroz)
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
# Various ways to summarize the data, which with its pros and cons
summary(Mroz)
```

```
##   lfp            k5              k618             age          wc         hc     
##  no :325   Min.   :0.0000   Min.   :0.000   Min.   :30.00   no :541   no :458  
##  yes:428   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:36.00   yes:212   yes:295  
##            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                      
##       lwg               inc        
##  Min.   :-2.0541   Min.   :-0.029  
##  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  
```

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

```
## Rows: 753
## Columns: 8
## $ lfp  <fct> yes, yes, yes, yes, yes, yes, yes, yes, yes, yes, yes, yes, ye...
## $ k5   <int> 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 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, 4, 2,...
## $ age  <int> 32, 30, 35, 34, 31, 54, 37, 54, 48, 39, 33, 42, 30, 43, 43, 35...
## $ wc   <fct> no, no, no, no, yes, no, yes, no, no, no, no, no, no, no, no, ...
## $ hc   <fct> no, no, no, no, no, no, no, no, no, no, no, yes, yes, no, yes,...
## $ lwg  <dbl> 1.2101647, 0.3285041, 1.5141279, 0.0921151, 1.5242802, 1.55648...
## $ inc  <dbl> 10.910001, 19.500000, 12.039999, 6.800000, 20.100000, 9.859000...
```

```r
#View(Mroz)
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
##
## lowest : 0 1 2 3 4, highest: 4 5 6 7 8
##
## 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
## ---------------------------------------------------------------------------
```

```
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
## 14  yes  0    2  43  no  no 0.8183691 14.60
## 155 yes  0    0  47 yes yes 2.1029148 32.70
## 280 yes  0    2  43 yes  no 1.3791769  8.56
## 377 yes  0    1  44 yes  no 0.5163968 14.52
## 553  no  2    2  30  no  no 0.8180865 35.00
```

```
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
```

Descriptive statistical analysis of the data

Breakout room discussion: Task: Discuss the basic descriptive data analysis below; feel free to add more analyses as you see fit.

An initiation of the exploratory data analysis (EDA):

- *Note that this descriptive statistics analysis I included here is far from completed, and you can use it as a practice to complete it. Feel free to work with your classmates.*

1. No variable in the data set has missnig value. (This is very unlikely in practice, but this is a clean dataset highly curated for used in this example.)
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. Because our focus on this example is logitic regression modeling, let's for the time being, pretend that this list is sufficient (that is, I completely assume away the issue of omitted variable bias.)

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. In this course, I pay a lot of attention to how students conduct EDA, much more so than you would in w203. (*I will tell you why it matters in practice.*)

In general, we will examine / 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

Note that for some of the graphs below, such as the overlapping density functions, I plotted them to show you their effectiveness, or lack thereof, in displaying the underlying relationship.

Note that unlike the async lectures, which I didn't use any specific libraries to conduct data visualization, I use *ggplot()* quite extensively in all of the live sessions.

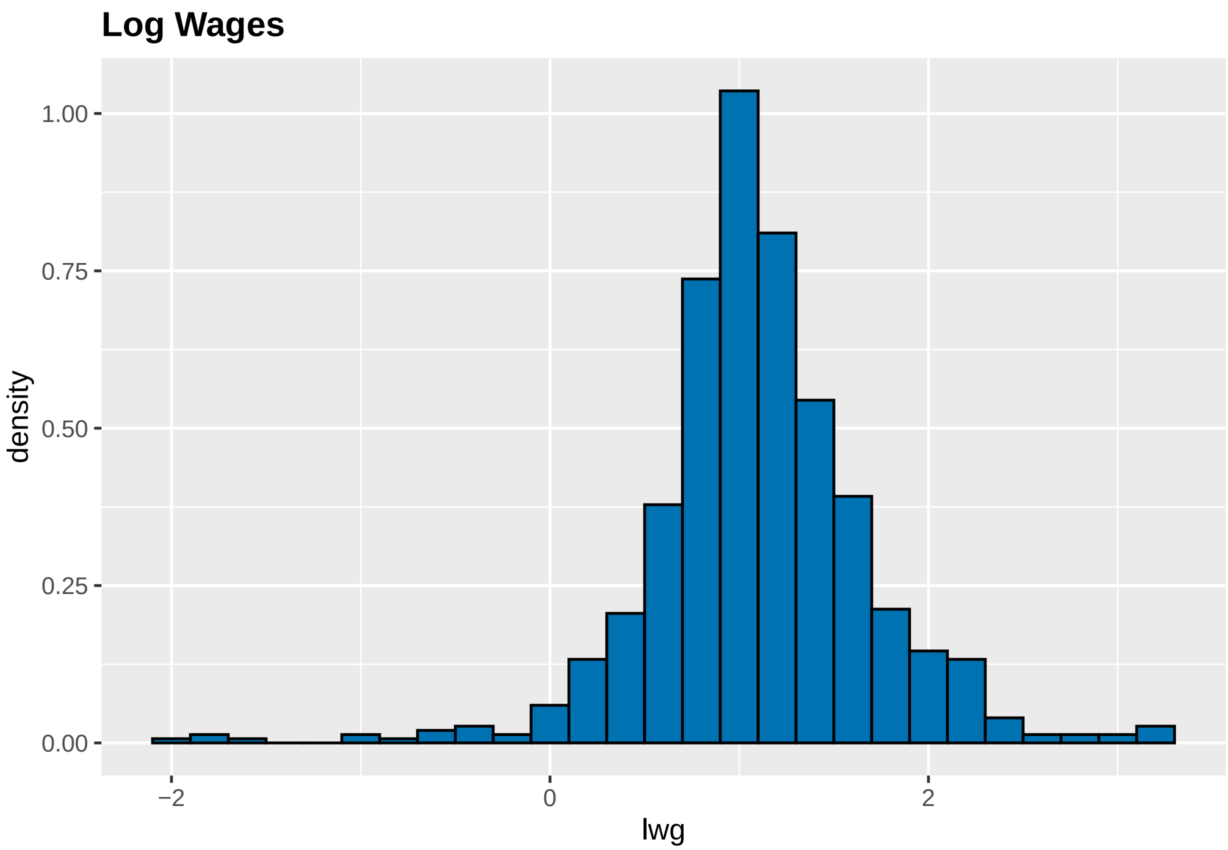
```
library(dplyr)
library(ggplot2)

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
```

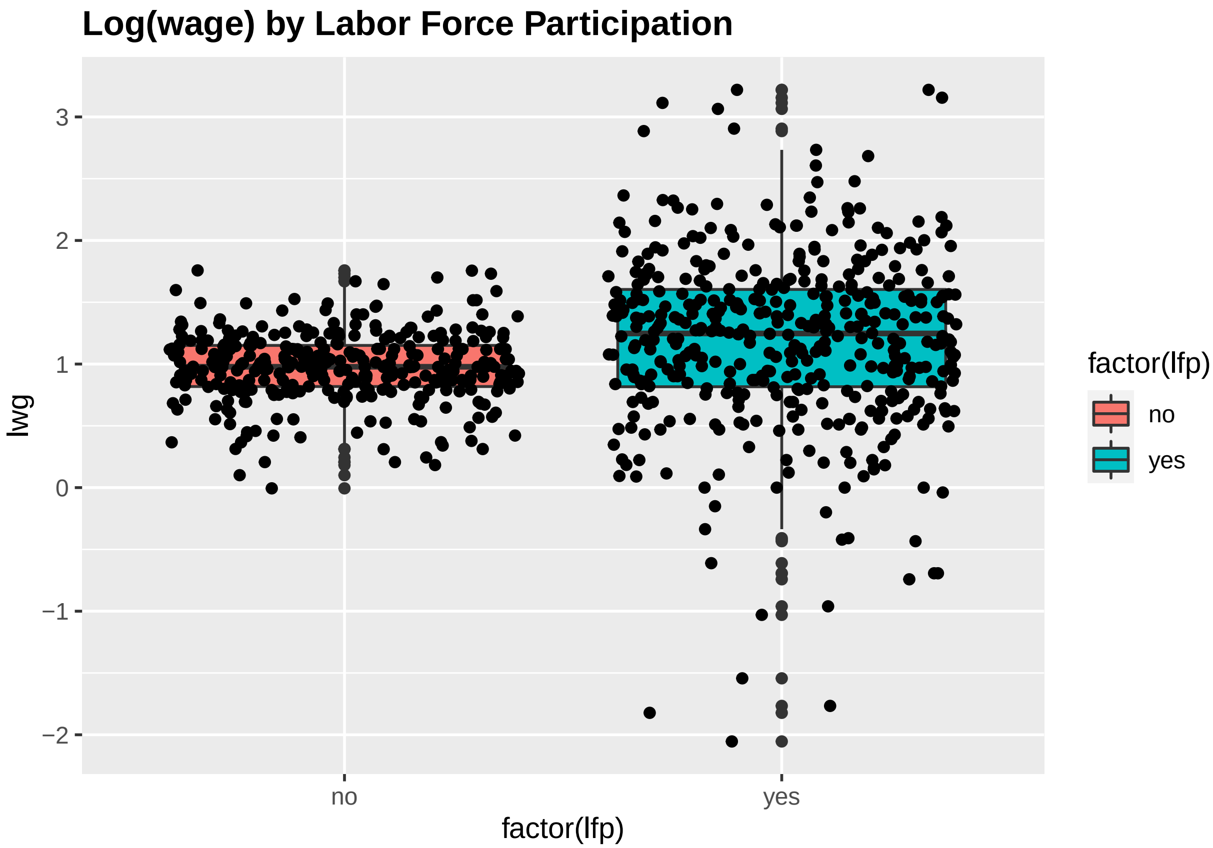
```
min(exp(Mroz$lwg))
```

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

```
# 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"))
```



```r
# 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"))
```

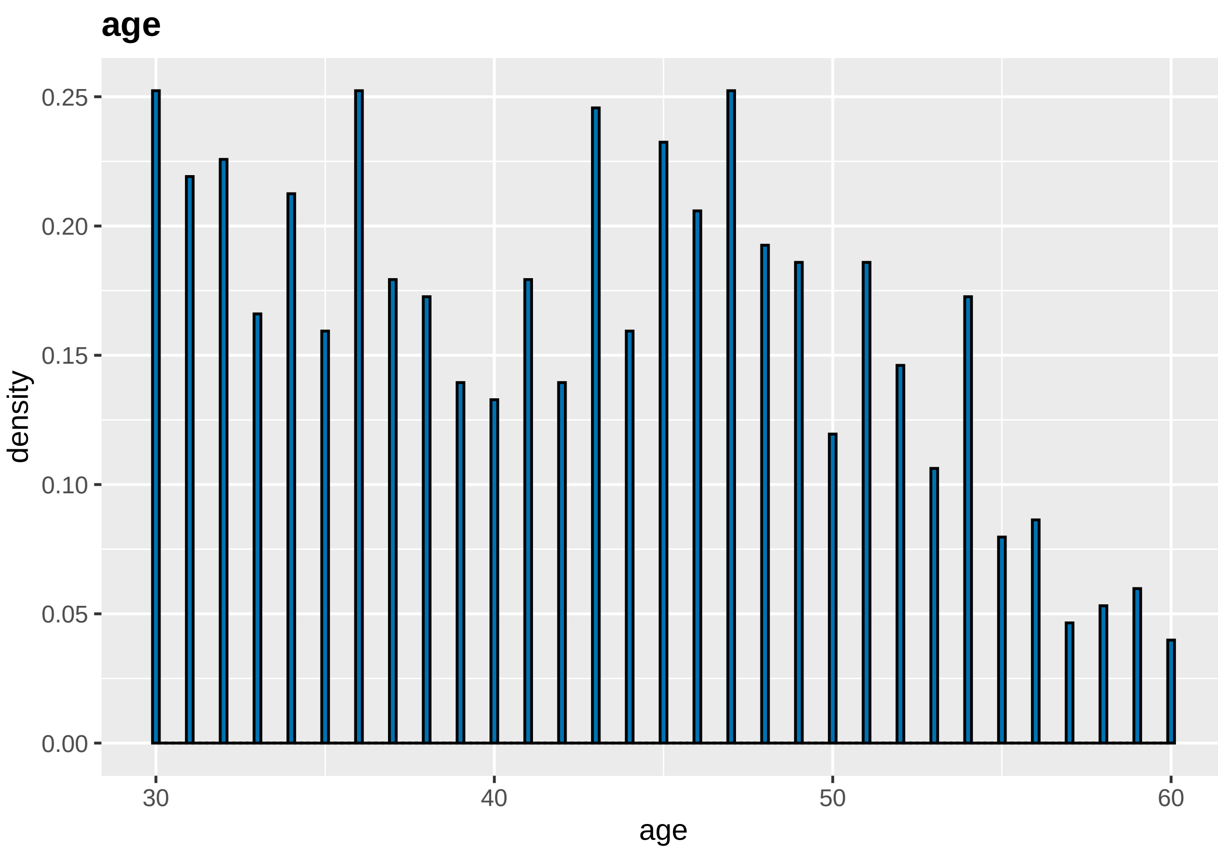



```r
# age by lfp
ggplot(Mroz, aes(factor(lfp), age)) +
  geom_boxplot(aes(fill = factor(lfp))) +
  geom_jitter() +
  ggtitle("Age by Labor Force Participation") +
  theme(plot.title = element_text(lineheight=1, face="bold"))
```

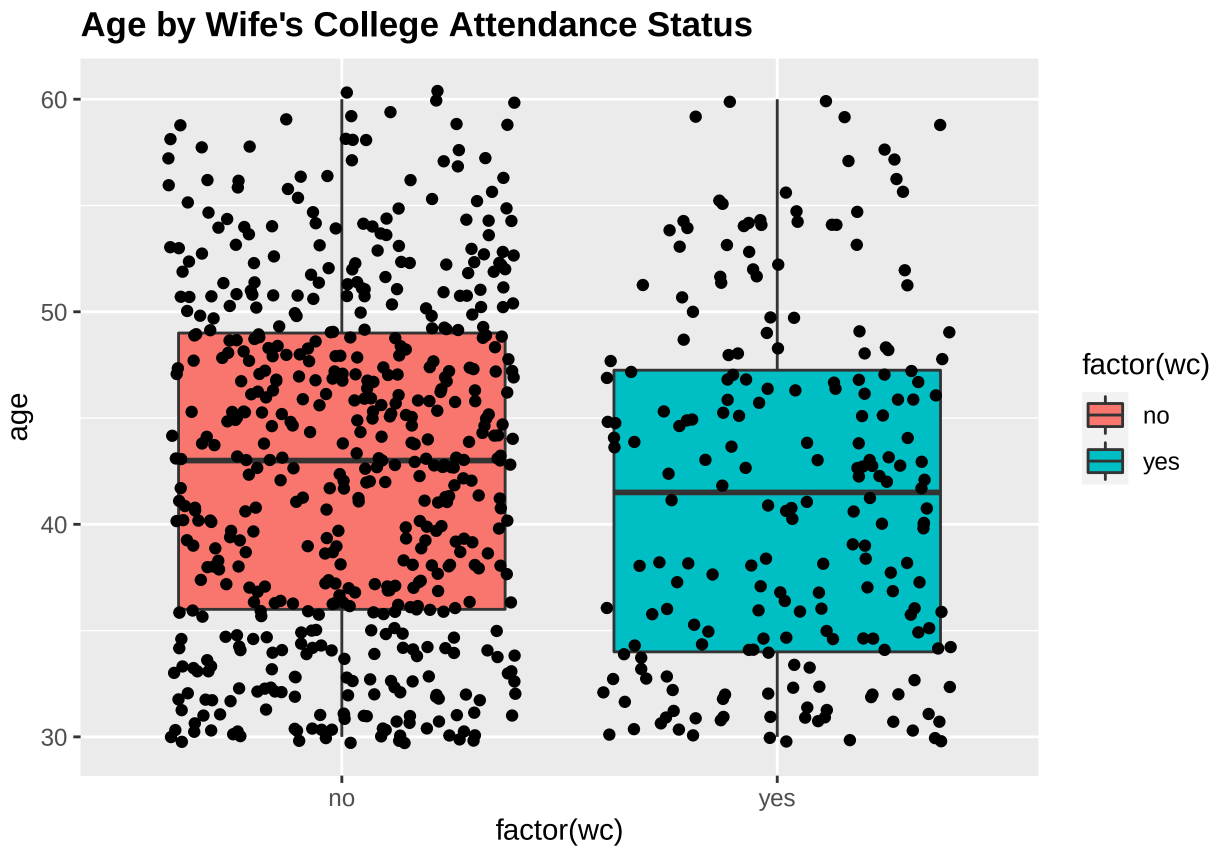

```
# Distribution of age
summary(Mroz$age)
```

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

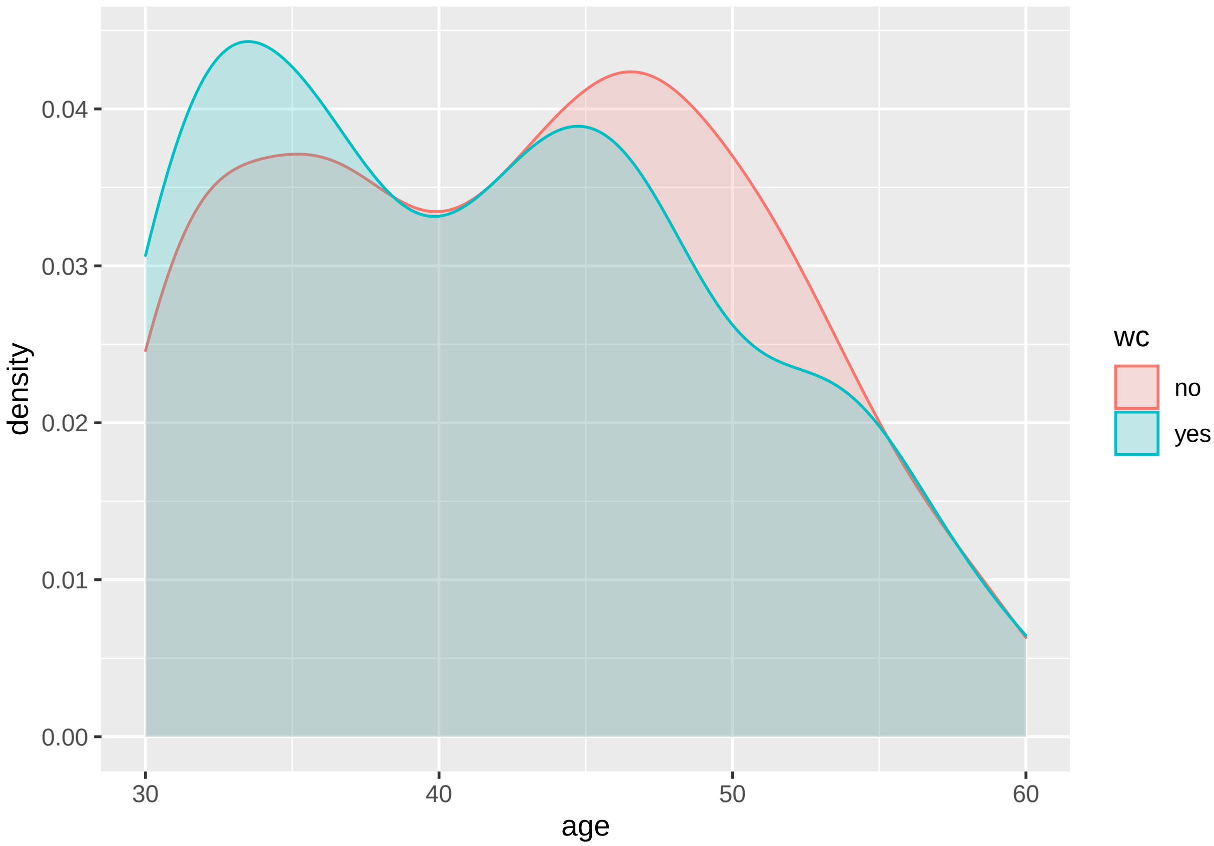
```
ggplot(Mroz, aes(x = age)) +
  geom_histogram(aes(y = ..density..), binwidth = 0.2, fill="#0072B2", colour="black") +
  ggtitle("age") +
  theme(plot.title = element_text(lineheight=1, face="bold"))
```



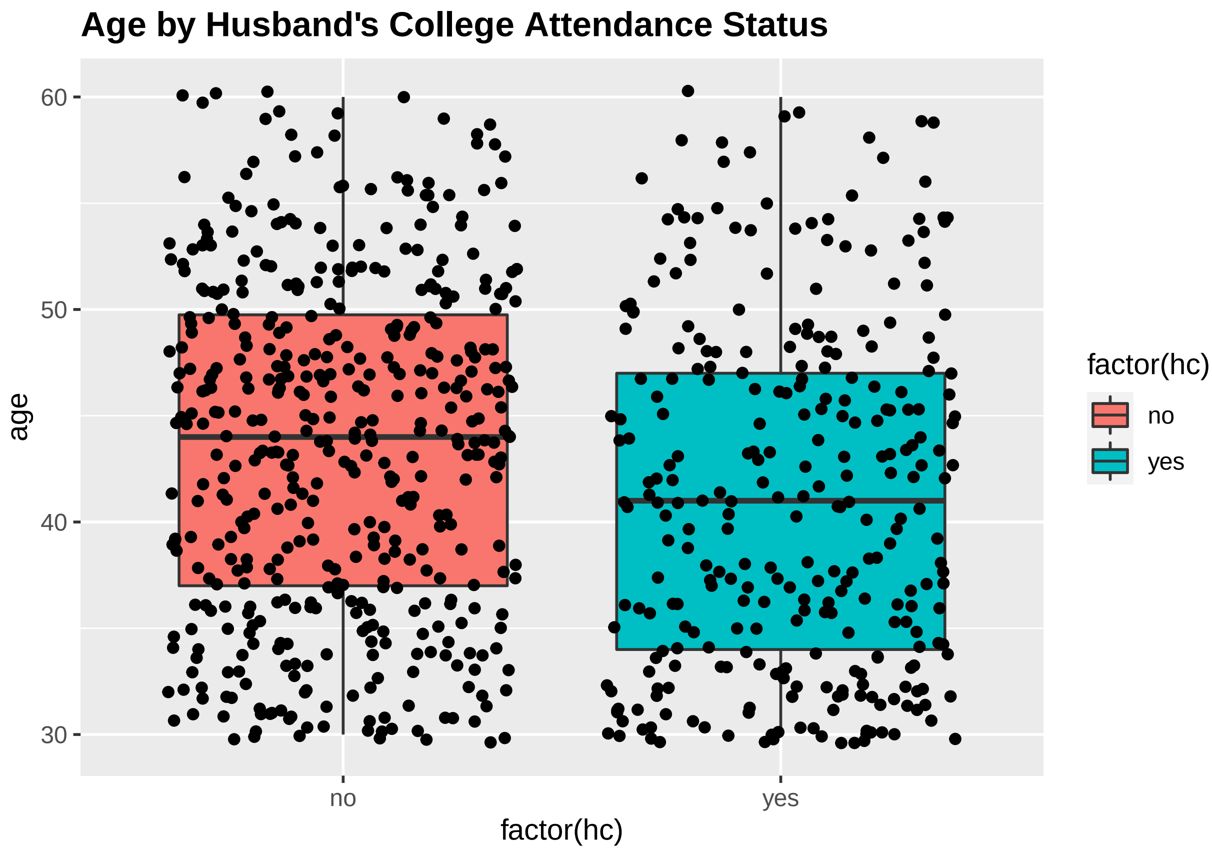

```
# Distribution of age by wc
# Were those who attended colleage tend to be younger?
ggplot(Mroz, aes(factor(wc), age)) +
  geom_boxplot(aes(fill = factor(wc))) +
  geom_jitter() +
  ggtitle("Age by Wife's College Attendance Status") +
  theme(plot.title = element_text(lineheight=1, face="bold"))
```



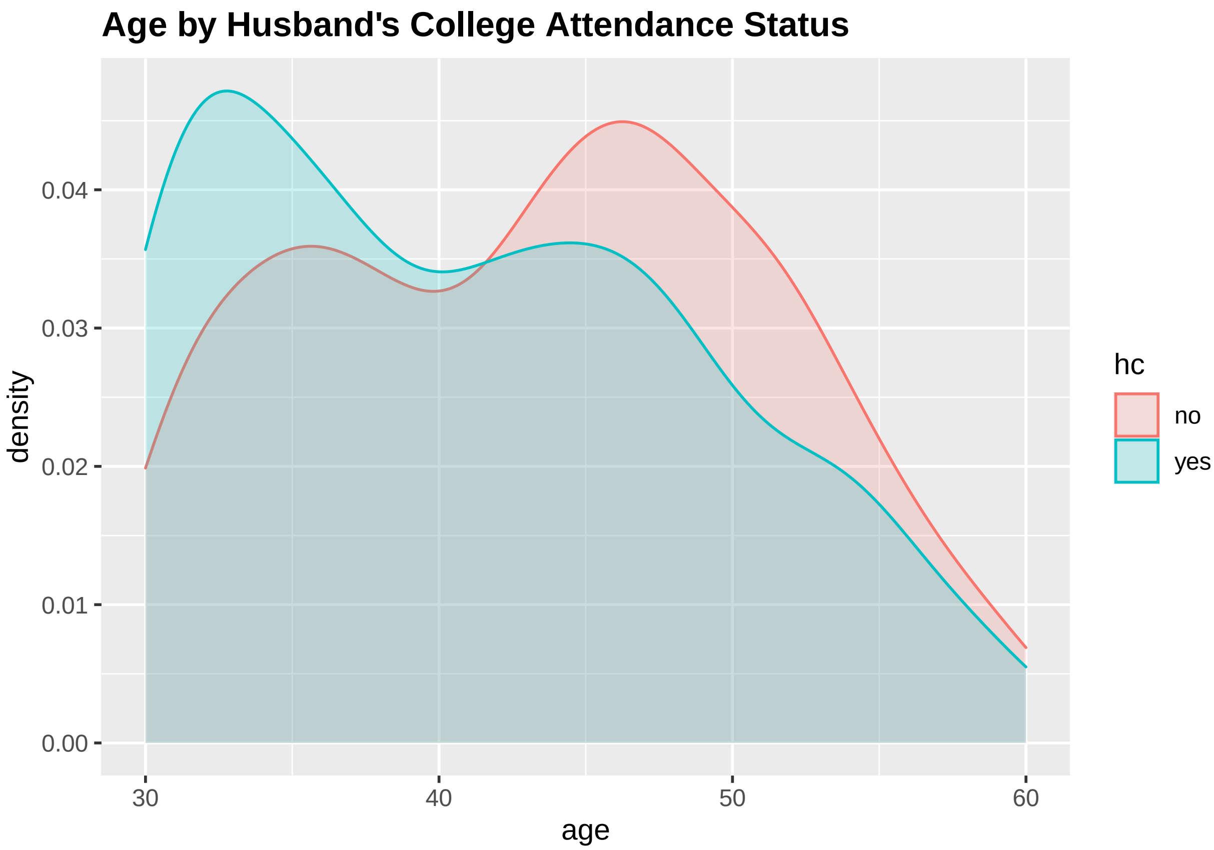

```
ggplot(Mroz, aes(age, fill = wc, colour = wc)) +
  geom_density(alpha=0.2)
```



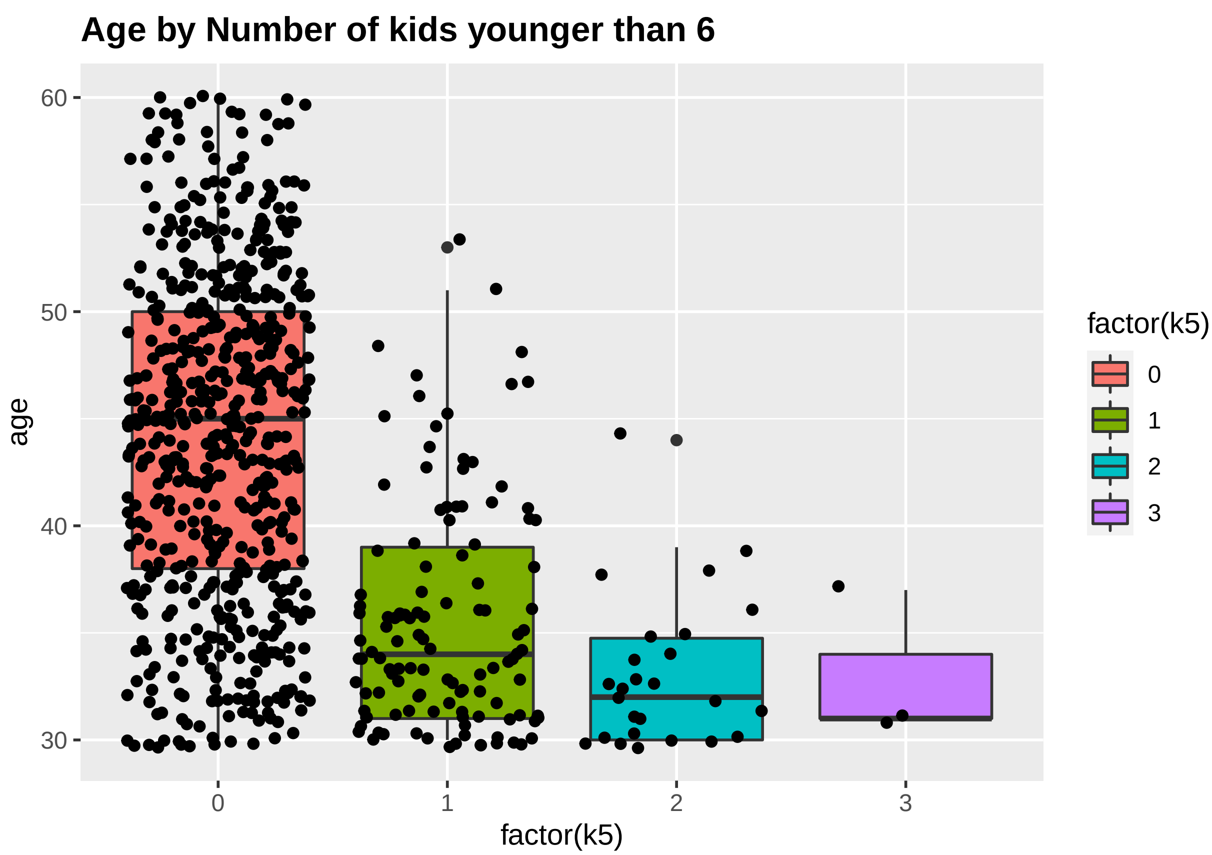

```r
# Distribution of age by hc
# Were those whose husband attended colleage tend to be younger?
ggplot(Mroz, aes(factor(hc), age)) +
  geom_boxplot(aes(fill = factor(hc))) +
  geom_jitter() +
  ggtitle("Age by Husband's College Attendance Status") +
  theme(plot.title = element_text(lineheight=1, face="bold"))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



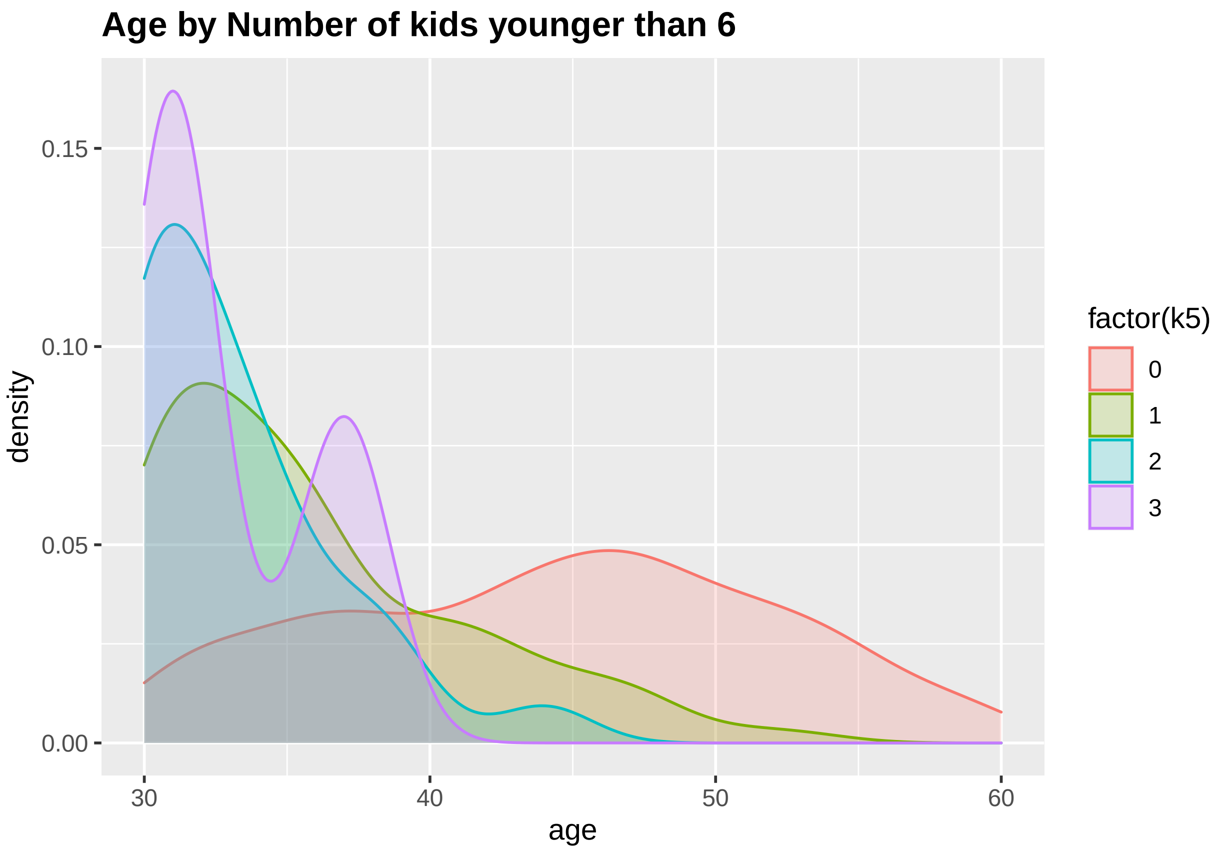

```
ggplot(Mroz, aes(age, fill = hc, colour = hc)) +
  geom_density(alpha=0.2) +
  ggtitle("Age by Husband's College Attendance Status") +
  theme(plot.title = element_text(lineheight=1, face="bold"))
```



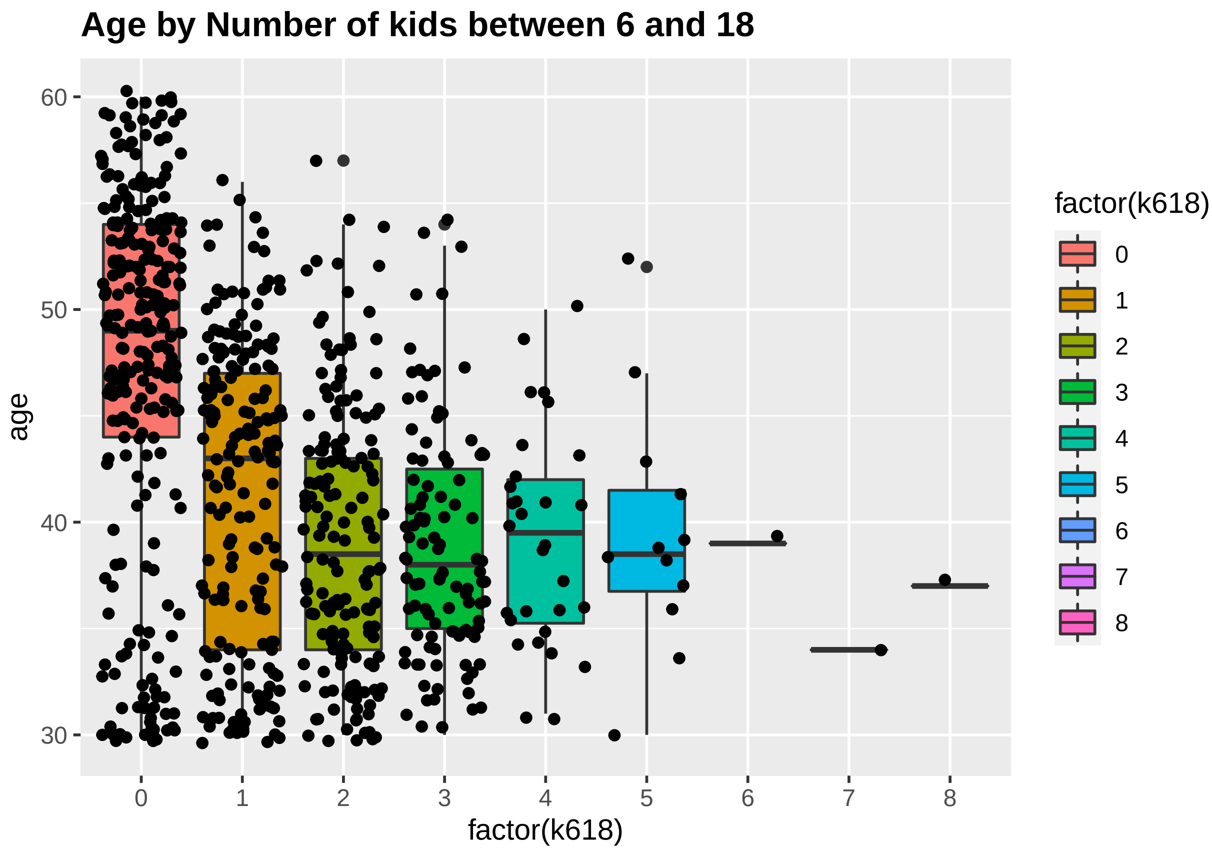

```
# Distribution of age by number kids in different age group
ggplot(Mroz, aes(factor(k5), age)) +
  geom_boxplot(aes(fill = factor(k5))) +
  geom_jitter() +
  ggtitle("Age by Number of kids younger than 6") +
  theme(plot.title = element_text(lineheight=1, face="bold"))
```




```
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"))
```

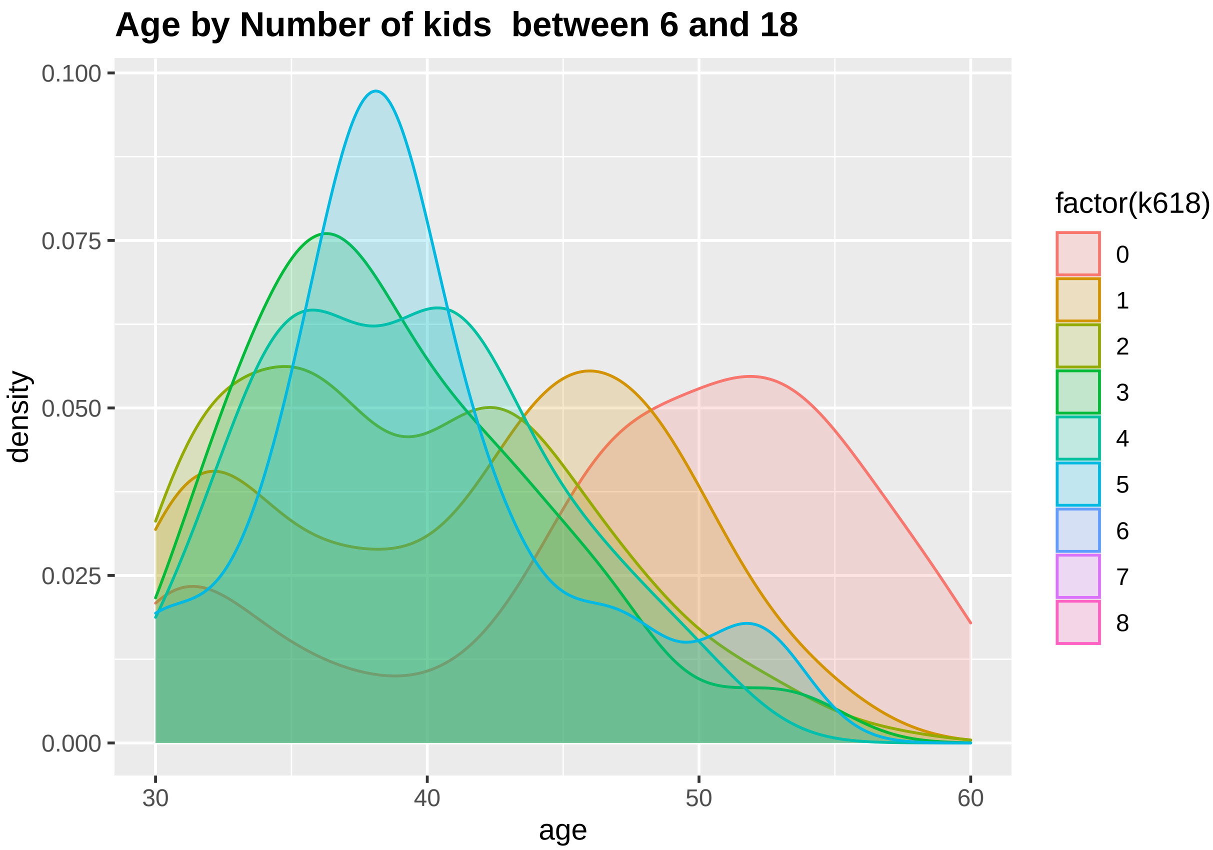



```
ggplot(Mroz, aes(factor(k618), age)) +
  geom_boxplot(aes(fill = factor(k618))) +
  geom_jitter() +
  ggtitle("Age by Number of kids between 6 and 18") +
  theme(plot.title = element_text(lineheight=1, face="bold"))
```




```
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"))
```

```
## Warning: Groups with fewer than two data points have been dropped.

## Warning: Groups with fewer than two data points have been dropped.

## Warning: Groups with fewer than two data points have been dropped.

## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning -
## Inf

## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning -
## Inf

## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning -
## Inf
```




```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
```

As a best practice, we will need to incorporate insights generated from EDA on model specification. In what follows, I employ a very simple specification that uses all the variables as-is, but the focus is on how to interpret the coefficients.

Estimate a Binary Logistic Regression

Again, I have not used any EDA to inform the specification of my model, something that I take very seriously about in this course. The reason is that we will be talking about various techniques of variable transformation for binary logistic regression next week, and I want to wait till next week to incorporate "insights" from EDA for model specification.

Breakout Room Discussion:

- **Ensure you understand the model estimation procedure and the model outputs**
- **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 ratios) of decreasing k5 by 1-unit.**

- **Interpret the effect (in terms of odds rations) of decreasing inc by $10,000.**
- **Discuss the result of the test.**

```r
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
```

```r
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
```

```r
vcov(mroz.glm)
```

```
##               (Intercept)            k5          k618           age
## (Intercept)  0.4152192592 -0.0630518516 -2.303486e-02 -7.666271e-03
## k5          -0.0630518516  0.0388092385  1.957324e-03  1.221579e-03
```

```
## k618        -0.0230348597  0.0019573238  4.624113e-03  3.747432e-04
## age         -0.0076662713  0.0012215794  3.747432e-04  1.634074e-04
## wcyes        0.0128187729 -0.0045497706  7.302961e-04 -1.276189e-04
## hcyes       -0.0124953266 -0.0028554298 -1.360980e-04  2.797675e-04
## lwg         -0.0188134789 -0.0009772917  7.584108e-04 -5.428161e-05
## inc         -0.0006091469  0.0001235370 -3.116678e-05 -8.380831e-06
##                     wcyes         hcyes           lwg           inc
## (Intercept)  0.0128187729 -0.0124953266 -1.881348e-02 -6.091469e-04
## k5          -0.0045497706 -0.0028554298 -9.772917e-04  1.235370e-04
## k618         0.0007302961 -0.0001360980  7.584108e-04 -3.116678e-05
## age         -0.0001276189  0.0002797675 -5.428161e-05 -8.380831e-06
## wcyes        0.0528907469 -0.0207304484 -6.736742e-03 -2.532608e-04
## hcyes       -0.0207304484  0.0424523656  1.434414e-04 -4.897312e-04
## lwg         -0.0067367419  0.0001434414  2.274594e-02 -1.077886e-04
## inc         -0.0002532608 -0.0004897312 -1.077886e-04  6.737744e-05
```

Interpretation of model results

Do the "raw" coefficient estimates directionally make sense?

```
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
```

Below, I include some codes to help you interpret the model results. Feel free to modify the codes.

Interpreting the coefficient estimates in terms of odds ratio is a common practice. Recall that

$$
\begin{aligned}
OR &= \frac{Odds_{x_k+c}}{Odds_{x_k}} \\
&= \frac{exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k(x_k + c) + \cdots + \beta_K x_K)}{exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k(x_k) + \cdots + \beta_K x_K} \\
&= \frac{exp(\beta_0)exp(\beta_1 x_1)\cdots exp(\beta_k(x_k + c))\cdots exp(\beta_K x_K)}{exp(\beta_0)exp(\beta_1 x_1)\cdots exp(\beta_k(x_k))\cdots exp(\beta_K x_K)} \\
&= \frac{exp(\beta_k(x_k + c))}{exp(\beta_k(x_k))} \\
&= exp\{\beta_k x_k + \beta_k c - \beta_k x_k)\} \\
&= exp(c\beta_k)
\end{aligned}
$$

The odds of a success change by $exp(c\beta_k)$ times for every *c-unit* increase in x_k.

Importantly, the change in the odds of a success is not a function of the $X's$!

It's also common to say "increase" instead of "change" when $exp(c\beta_k) > 1$ and "decrease" when $exp(c\beta_k) < 1$.

The estimated odds ratio becomes

$$\widehat{OR} = \frac{Odds_{x_k+c}}{Odds_{x_k}} = exp(c\hat{\beta}_k)$$

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

```
##              [,1]
## (Intercept) 24.10
## k5           0.23
## k618         0.94
## age          0.94
## wcyes        2.24
## hcyes        1.12
## lwg          1.83
## inc          0.97
```

```
#c = YOU NEED TO SPECIFY THE NUMBER HERE
c=-1
exp(c*coef(mroz.glm)['inc'])
```

```
##      inc
## 1.035047
```

<You should interpret The odds of participating in the labor force change.>

```
#c = YOU NEED TO SPECIFY THE NUMBER HERE
c=-1
exp(c*coef(mroz.glm)['k5'])
```

```
##       k5
## 4.318521
```

<You should interpret The odds of participating in the labor force change.>

Statistical Inference

Breakout Room Discussion (10 minutes):

- **Discuss the results of the test.**

Using Likelihood Ratio Test (LRT) for hypothesis testing, such as, in a logistic regression model, $logit(\pi) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \cdots + \beta_K x_K$, test

$H_0 : \beta_k = 0$ $H_a : \beta_k \neq 0$

For instance, suppose we want to test whether family income (*inc*) has an effect on the wife's labor force participation, we test

$H_0 : \beta_{inc} = 0$ $H_a : \beta_{inc} \neq 0$

Using LRT, implemented via the *Anova()* (or *anova()*) function.

$$
\begin{aligned}
-2log(\Lambda) &= -2log\left(\frac{L(\hat{\beta}^{(0)}|y_1,\ldots,y_n)}{L(\hat{\beta}^{(a)}|y_1,\ldots,y_n)}\right) \\
&= -2\sum y_i log\left(\frac{\hat{\pi}_i^{(0)}}{\hat{\pi}_i^{(a)}}\right) + (1-y_i)log\left(\frac{1-\hat{\pi}_i^{(0)}}{1-\hat{\pi}_i^{(a)}}\right)
\end{aligned}
$$

```r
# Likelihood Ratio Test
library(car)
Anova(mroz.glm, test="LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: lfp
##      LR Chisq Df Pr(>Chisq)
## k5     66.484  1  3.527e-16 ***
## k618    0.903  1   0.342042
## age    25.598  1  4.204e-07 ***
## wc     12.724  1   0.000361 ***
## hc      0.294  1   0.587489
## lwg    17.001  1  3.736e-05 ***
## inc    19.504  1  1.004e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that another way to perform hypothesis testing is to use *anova()* function to estimate both models under the null hypothesis and alternative hypothesis and then use the corresponding model-fitted objects as argument within the function. This is my preferred method. As an illustration, examine the following example.

```r
mroz.glm.h0 <- glm(lfp ~ k5 + k618 + age + wc + hc + lwg,
               family = binomial, data = Mroz)
mroz.glm.h1 <- glm(lfp ~ k5 + k618 + age + wc + hc + lwg + inc,
               family = binomial, data = Mroz)
anova(mroz.glm.h0,mroz.glm.h1)
```

```
## Analysis of Deviance Table
##
## Model 1: lfp ~ k5 + k618 + age + wc + hc + lwg
## Model 2: lfp ~ k5 + k618 + age + wc + hc + lwg + inc
##   Resid. Df Resid. Dev Df Deviance
## 1       746     924.77
## 2       745     905.27  1   19.504
```

Confidence Interval for β_k

Wald Confidence:

$$\hat{\beta}_k \pm Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{\beta}_k)}$$

$$exp\left(\hat{\beta}_k \pm Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{\beta}_k)}\right)$$

However, for reasons we discussed extensively in lecture 1, Wald confidence interval only has true confidence level close to the stated confidence level when the sample is sufficiently large. Therefore, we use the *profile likelihood ratio (LR)* confidence interval, which, for binary logistic regression, can be calculated using a R function $confint()$:

```
#round(exp(cbind(Estimate=coef(mroz.glm), confint(mroz.glm))),2)
confint.default(object=mroz.glm, level=0.95)
```

```
##                   2.5 %      97.5 %
## (Intercept)  1.91918849  4.44509244
## k5          -1.84902713 -1.07679895
## k618        -0.19784986  0.06870849
## age         -0.08792495 -0.03781615
## wcyes        0.35652149  1.25802607
## hcyes       -0.29209685  0.51556400
## lwg          0.30909613  0.90029012
## inc         -0.05053455 -0.01835831
```

```
exp(confint.default(object=mroz.glm, level=0.95))
```

```
##                 2.5 %     97.5 %
## (Intercept) 6.8154254 85.2077537
## k5          0.1573902  0.3406843
## k618        0.8204930  1.0711239
## age         0.9158296  0.9628899
## wcyes       1.4283522  3.5184694
## hcyes       0.7466962  1.6745827
## lwg         1.3621933  2.4603168
## inc         0.9507211  0.9818092
```

Wald Confidence Interval

```
#vcov(mroz.glm)
#summary(mroz.glm)
mroz.glm$coefficients[8] + qnorm(p = c(0.025, 0.975))*sqrt(vcov(mroz.glm)[8,8])
```

```
## [1] -0.05053455 -0.01835831
```

```
exp(mroz.glm$coefficients[8] + qnorm(p = c(0.025, 0.975))*sqrt(vcov(mroz.glm)[8,8]))
```

```
## [1] 0.9507211 0.9818092
```

Confidence Interval for the Probability of Success

Recall that the estimated probability of success is

$$\hat{\pi} = \frac{exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_K x_k\right)}{1 + exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_K x_k\right)}$$

While backing out the estimated probability of success is straight-forward, obtaining its confidence interval is not, as it involves many parameters.

Wald Confidence Interval

$$\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_K x_K \pm Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_K x_K)}$$

where

$$\widehat{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_K x_K) = \sum_{i=0}^{K} x_i^2 \widehat{Var}(\hat{\beta}_i) + 2\sum_{i=0}^{K-1}\sum_{j=i+1}^{K} x_i x_j \widehat{Cov}(\hat{\beta}_i, \hat{\beta}_j)$$

So, the Wald Interval for π

$$\frac{exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_K x_k \pm \sqrt{\sum_{i=0}^{K} x_i^2 \widehat{Var}(\hat{\beta}_i) + 2\sum_{i=0}^{K-1}\sum_{j=i+1}^{K} x_i x_j \widehat{Cov}(\hat{\beta}_i, \hat{\beta}_j)}\right)}{1 + exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_K x_k\right) \pm \sqrt{\sum_{i=0}^{K} x_i^2 \widehat{Var}(\hat{\beta}_i) + 2\sum_{i=0}^{K-1}\sum_{j=i+1}^{K} x_i x_j \widehat{Cov}(\hat{\beta}_i, \hat{\beta}_j)}}$$

```r
alpha = 0.5

wc = "yes"
hc = "yes"
predict.data <- data.frame(k5 = mean(Mroz$k5),
                           k618 = mean(Mroz$k618),
                           age = mean(Mroz$age),
                           wc = factor(wc),
                           hc = factor(hc),
                           lwg = mean(Mroz$lwg),
                           inc = mean(Mroz$inc))
str(predict.data)
```

```
## 'data.frame':    1 obs. of  7 variables:
##  $ k5  : num 0.238
##  $ k618: num 1.35
##  $ age : num 42.5
##  $ wc  : Factor w/ 1 level "yes": 1
##  $ hc  : Factor w/ 1 level "yes": 1
##  $ lwg : num 1.1
##  $ inc : num 20.1
```

```r
# Obtain the linear predictor
linear.pred = predict(object = mroz.glm, newdata = predict.data,
                      type = "link", se = TRUE)
linear.pred
```

```
## $fit
##         1
```

```
## 0.9616785
##
## $se.fit
## [1] 0.1823138
##
## $residual.scale
## [1] 1
```

```
# Then, compute pi.hat
pi.hat = exp(linear.pred$fit)/(1+exp(linear.pred$fit))
pi.hat
```

```
##         1
## 0.7234578
```

```
# Compute Wald Confidence Interval (in 2 steps)
# Step 1
CI.lin.pred = linear.pred$fit + qnorm(p = c(alpha/2, 1-alpha/2))*linear.pred$se
CI.lin.pred
```

```
## [1] 0.8387098 1.0846473
```

```
# Step 2
CI.pi = exp(CI.lin.pred)/(1+exp(CI.lin.pred))
CI.pi
```

```
## [1] 0.6981934 0.7473724
```

```
# Store all the components in a data frame
str(predict.data)
```

```
## 'data.frame':    1 obs. of  7 variables:
##  $ k5  : num 0.238
##  $ k618: num 1.35
##  $ age : num 42.5
##  $ wc  : Factor w/ 1 level "yes": 1
##  $ hc  : Factor w/ 1 level "yes": 1
##  $ lwg : num 1.1
##  $ inc : num 20.1
```

```
round(data.frame(pi.hat, lower=CI.pi[1], upper=CI.pi[1]),4)
```

```
##   pi.hat  lower  upper
## 1 0.7235 0.6982 0.6982
```

Visualize 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
```

```r
summary(Mroz)
```

```
##    lfp            k5              k618             age            wc         hc     
##  no :325   Min.   :0.0000   Min.   :0.000   Min.   :30.00   no :541   no :458  
##  yes:428   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:36.00   yes:212   yes:295  
##            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                        
##       lwg               inc        
##  Min.   :-2.0541   Min.   :-0.029  
##  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
```

```r
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
```

```r
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" ...
```

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

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

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

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

```r
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
```

```
# Effect of income on LFP for a family with no kid, wife was 40 years old, both wife and husband attend

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
```

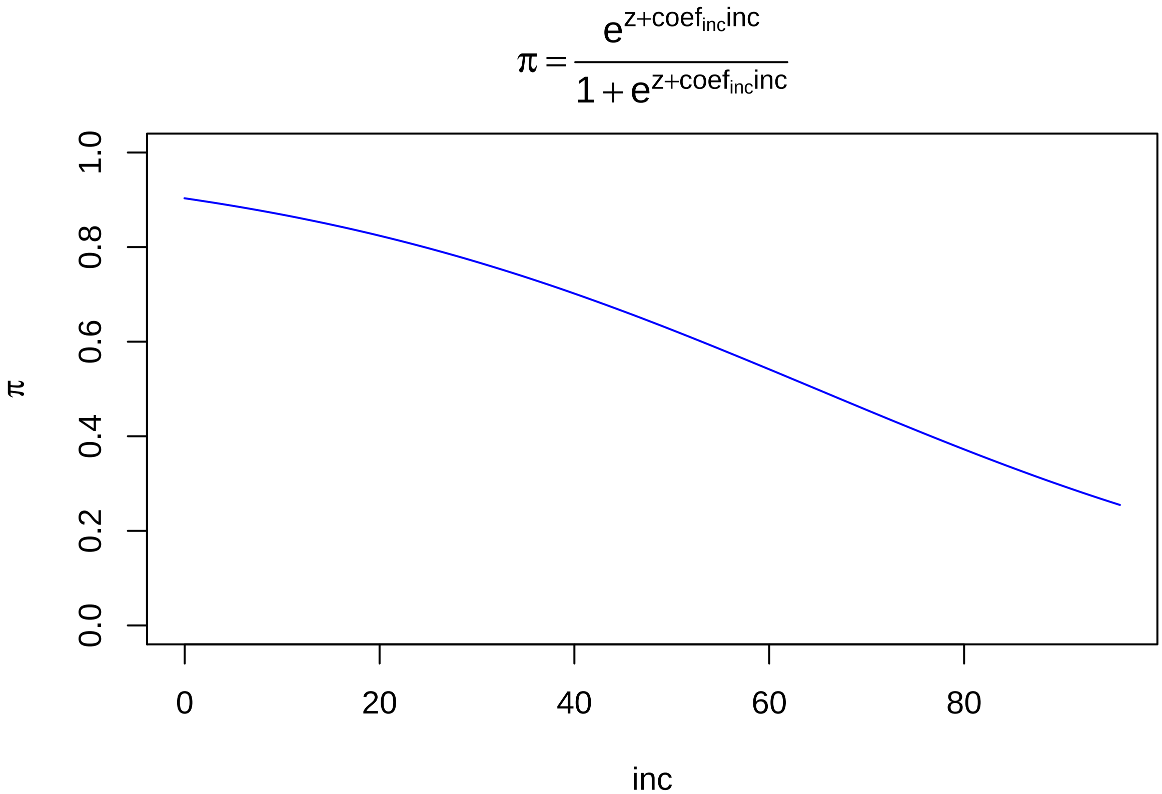
```
z = coef[1]*xx[1] + coef[2]*xx[2] + coef[3]*xx[3] + coef[3]*xx[3] + coef[4]*xx[4] + coef[5]*xx[5] + coe
z
```

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

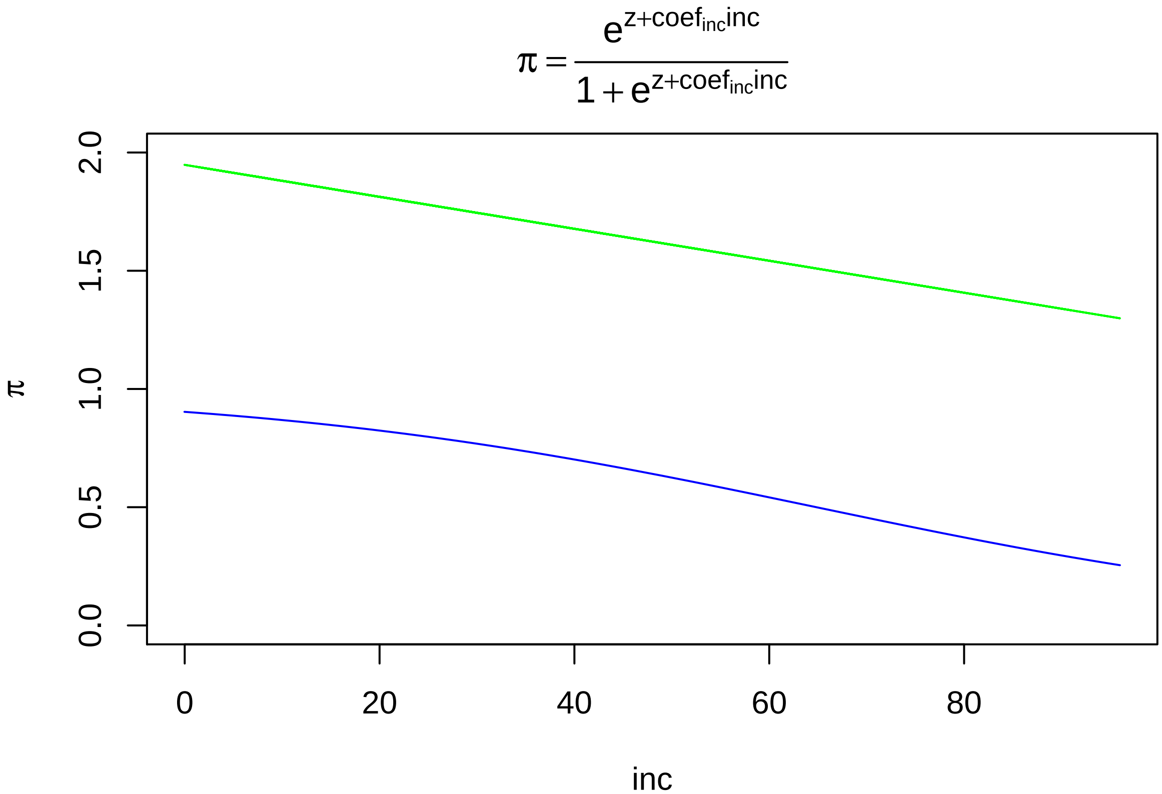
```
x <- Mroz$inc
coef[8]
```

```
##         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))
```




```
# 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] + l

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


$\pi = \frac{e^{z+coef_{inc}inc}}{1+e^{z+coef_{inc}inc}}$
2.0
1.5
1.0
0.5
0.0
π
0
20
40
60
80
inc