Assignment 8

Mingrui Duan

Preparation for the data

Question 1: To visualize this data we'll bin age. The binning on age is only for visualization and will allow you to put a single point with error bars for each bin. List advantages and disadvantages of binning by (a) year, (b) month, (c) quantiles (e.g. 10% quantiles).

1. Binning by "year":

    1) Advantages: Sorted in year can be used to find common features from the dataset.
    2) Disadvantages: The sample size can be small, which will decrease the accuracy of the model; Also, binned by "year" can make the bin too broad, so it may lose some information.

2. Binning by "month":

    1) Advantages: The amount of data points is large, the regression model can be more accurate.
    2) Disadvantages Different to find common traits among these samples; While the dataset size is too big, analysing work can be extremely heavy and have high computation cost.

3. Binning by "quantiles":

    1) Advantages: For complex data, binned by "quantiles" can simplify the data and easy to interpret the findings; It will also reduce sensitivity to the outliers.
    2) Disadvantages: Binning data by "quantiles" can result in ignorance of individual values, focusing too much on the distribution of data.

Question 2: (a) Choose the best option for binning and plot the results. Make this a "publication quality" plot: include error bars, points for the mean, axis labels, and faceting by language. (b) Are the best error bars from mean_se or mean_cl_boot? (c) Finally, include in this plot a stat_smooth using a glm (NOT a lm). You may need to read online to see how to do this. Ensure that the plot goes from 0 to 12 and that the stat summary line extends through this range. (d) What does the plot show? What basic patterns should you expect to see in a logistic regression based on this plot?

```
library("ggplot2")
library("dplyr")
```

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

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

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

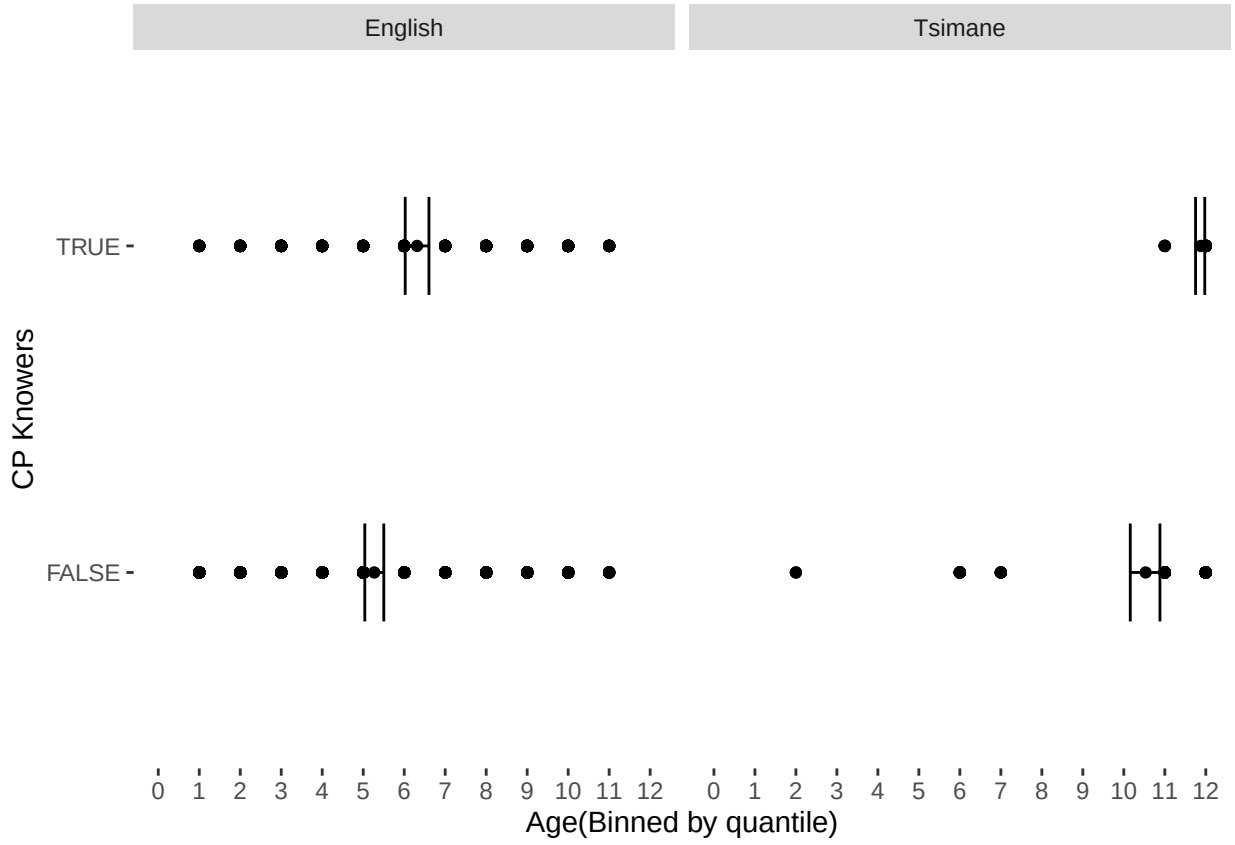
```
# Bin the data by year & Get the mean and sd of data after binned
d <- d %>%
   dplyr::mutate(AgeByQuantile = ntile(Age, 12))

d_summary <- d %>%
  dplyr::group_by(AgeByQuantile, Language) %>%
  dplyr::summarize(CP_mean = mean(IsCPKnowers), CP_sd = sd(IsCPKnowers))
```

```
## `summarise()` has grouped output by 'AgeByQuantile'. You can override using the
## `.groups` argument.
```

```
ggplot(d, aes(x=AgeByQuantile, y=IsCPKnowers)) +
  # geom_bar(aes(x=AgeByQuantile, y=CP_mean), stat='identity') +
  stat_summary(fun.data = mean_cl_boot, geom = "errorbar", width=.3) +
  stat_summary(fun.y = mean, geom = "point") +
  geom_point() +
  stat_smooth(method = "glm", method.args = list(family = binomial)) +
  facet_grid(~Language) +
  xlim(0, 12) +
  scale_x_continuous(limits = c(0, 12), breaks = seq(0, 12, by = 1)) +
  xlab("Age(Binned by quantile)") +
  ylab("CP Knowers") +
  theme(panel.background = element_rect(fill = "transparent"),
        plot.background = element_rect(fill = "transparent", color = NA),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank())
```

```
## Warning: The `fun.y` argument of `stat_summary()` is deprecated as of ggplot2 3.3.0.
```

```
## i Please use the ‘fun‘ argument instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.

## Scale for x is already present.
## Adding another scale for x, which will replace the existing scale.

## ‘geom_smooth()‘ using formula = ’y ~ x’

## Warning: glm.fit: algorithm did not converge

## Warning: Computation failed in ‘stat_smooth()‘
## Caused by error:
## ! y values must be 0 <= y <= 1

## Warning: glm.fit: algorithm did not converge

## Warning: Computation failed in ‘stat_smooth()‘
## Caused by error:
## ! y values must be 0 <= y <= 1
```



b) The best error bar from “mean_cl_boot”. Because for this problem we are working on, we need to determine how “confident” we are in telling the probability that under certain “Age” of being a “CP Knower”.

c)

d) The plot shows that when the language is “English”, the learning time to be a “CP Knower” is much less than “Tsimane”. Also, the overall number of being a “CP Knower” is more. I expect to see there are independent variables to determine the outcome and shows different trend(relation) in different value.

Question 3: (a) Run a logistic regression Language*Age (don't use binned age in the regression) and write up the results as you would for a paper. (b) When you do this, write a short description of the figure like you might find in a paper ("Figure 1 shows...") and then present/interpret the regression results. (c) Is there a different age slope in the two languages? (d) Is there a different intercept? What might those mean?

a)

```
# Covert the "Language" column to 0/1
# When it's "English", it would be 1
d$CPKnower.bin <- ifelse(d$IsCPKnowers == TRUE, 1, 0)
d$IsEnglish <- is.element(d$Language, "English")
# Set the logistic function
inv_logit <- function (z) { 1/(1+exp(-z)) }
g <- glm( CPKnower.bin ~ Age*IsEnglish + IsEnglish + Age + 1, data = d, family = "binomial")
summary(g)
```

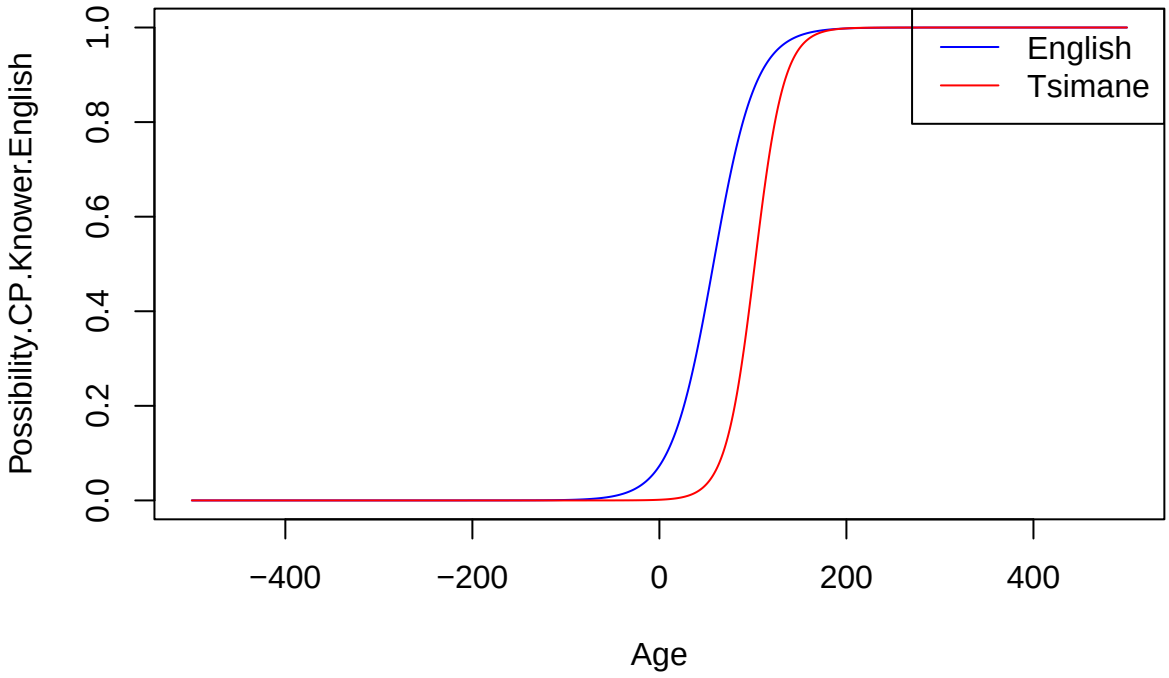
```
##
## Call:
## glm(formula = CPKnower.bin ~ Age * IsEnglish + IsEnglish + Age +
##     1, family = "binomial", data = d)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0325  -0.9797  -0.7370   1.2677   2.3563
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -6.57967    1.04144  -6.318 2.65e-10 ***
## Age                0.06447    0.01212   5.319 1.04e-07 ***
## IsEnglishTRUE      4.03882    1.10536   3.654 0.000258 ***
## Age:IsEnglishTRUE -0.02048    0.01443  -1.420 0.155626
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1537.8  on 1189  degrees of freedom
## Residual deviance: 1441.3  on 1186  degrees of freedom
## AIC: 1449.3
##
## Number of Fisher Scoring iterations: 5
```

So we conducted the logistic regression project Age to Language. The outcome is the possibility of an "CP knower" Speaker under given Age and Language they speak.

b) Figure 1 shows the relation between "Age" and the possibility of an "CP Knower" under certain Language they speak, where the y is more close to 1, means it's more likely to be a CP Knower under given Language.

```
Age <- seq(-500, +500, 0.1)
```

```r
Possibility.CP.Knower.English <- inv_logit(-6.57967 + 4.03882 + (0.06447-0.02048)*Age)
Possibility.CP.Knower.Tsimane <- inv_logit(-6.57967 + 0.06447*Age)

plot(Age, Possibility.CP.Knower.English, type = "l", col = "blue")
lines(Age, Possibility.CP.Knower.Tsimane, col = "red")
legend("topright", legend = c("English", "Tsimane"), col = c("blue", "red"), lty = 1)
```



c) There is a different slope.

d) There is a different intercept. It means that when the language they speak is different, the possibility to be a "CP Knower" under the same Age is also different.

Question 4: (a) Write 2-3 sentence as you might in a paper explaining the size of the Age coefficient (b) Using the regression coefficients, figure out the point at which Tsimane kids are 75% CPknowers vs. US kids and report this age. (c) What percent of Tsimane newborns are expected to be CPknowers according to the regression model, and does this number make sense? (d) At what age will exactly 100% of Tsimane kids be CP-knowers and does this number make sense?

(a): The size of Age coefficient is determined after binning the "Age". But When viewing these coefficients, the intercept value should also be included because there's baseline setting in regression.

(b):

```
age.pred.75.T <- (log(0.75/0.25) + 6.57967)/0.06447
age.pred.75.U <- (log(0.75/0.25) + 6.57967 - 4.03882)/(0.06447-0.02048)
```

According to the logistic function we just got, it says that when "Age" is 119.0985, Tsimane kids are 75% CP-Knowers. US kids are 82.73386.

(c):

```
prob.age.0 <- inv_logit(-6.57967)
```

It's 0.001386. It doesn't make sense. Because newborns can never be a CP-Knower.

(d): At the Age of positive infinity that 100% of Tsimane kids be CP-knowers. It's not make sense because nobody can live up to infinity.

Question 5: (a) Print a summary of both this and the dummy coded regression. (b) When you do this coding, are the coefficients added or subtracted for English (you should be able to tell by comparing the this to the previous regression)? (c) When you run this new regression, what happens to the age slope relative to the previous regression and why? (d) What happens to the language effect and interaction and why? (e) What happens to the intercept and why?

```
# Summary of the "dummy coding"
summary(g)
```

```
##
## Call:
## glm(formula = CPKnower.bin ~ Age * IsEnglish + IsEnglish + Age +
##     1, family = "binomial", data = d)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0325  -0.9797  -0.7370   1.2677   2.3563
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -6.57967    1.04144  -6.318 2.65e-10 ***
## Age                0.06447    0.01212   5.319 1.04e-07 ***
## IsEnglishTRUE      4.03882    1.10536   3.654 0.000258 ***
## Age:IsEnglishTRUE -0.02048    0.01443  -1.420 0.155626
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1537.8  on 1189  degrees of freedom
## Residual deviance: 1441.3  on 1186  degrees of freedom
## AIC: 1449.3
##
## Number of Fisher Scoring iterations: 5
```

```
g.1 <- glm( CPKnower.bin ~ Age*IsEnglish + IsEnglish + Age + 1,
            data = d,
            family = "binomial",
            contrasts = list(IsEnglish=contr.sum))
# Summary of the new coding scheme
summary(g.1)
```

```
##
## Call:
## glm(formula = CPKnower.bin ~ Age * IsEnglish + IsEnglish + Age +
##     1, family = "binomial", data = d, contrasts = list(IsEnglish = contr.sum))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0325  -0.9797  -0.7370   1.2677   2.3563
##
## Coefficients:
```

```
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -4.560261   0.552679  -8.251  < 2e-16 ***
## Age             0.054225   0.007213   7.518 5.56e-14 ***
## IsEnglish1     -2.019411   0.552679  -3.654 0.000258 ***
## Age:IsEnglish1  0.010241   0.007213   1.420 0.155626
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1537.8  on 1189  degrees of freedom
## Residual deviance: 1441.3  on 1186  degrees of freedom
## AIC: 1449.3
##
## Number of Fisher Scoring iterations: 5
```

(b) From the comparison, we can tell that the coefficients are added.

(c) The slope is much "neutral" compared to the previous one. This is mainly due to the reason that we consider both effect and choose no baseline (or create a new baseline which consider both sides effect).

(d) The language effect and interaction get weaker than the previous one. Because the new coding scheme mixed both side.

(e) The intercept become more neutral because these "sum coding" scheme. It's because we consider both effect instead of consider one effect first and adjust to the other on the basis of the "baseline". So the new intercept are more likely to be the middle. In sum coding scheme, the interaction effect is the difference between the average of one level and the average of another level of a categorical variable. In dummy coding scheme, the interaction effect is the product of two dummy variables representing two levels of a categorical variable.

Question 6: (a) Make the same plot as Q2 for Tsimane only but, now with grouping and color by Gender. (b) What do you see? What would you expect to see in a logistic regression? (c) Run the regression and report the results as you would in a paper, also talking through the figure "Figure 2 shows..."

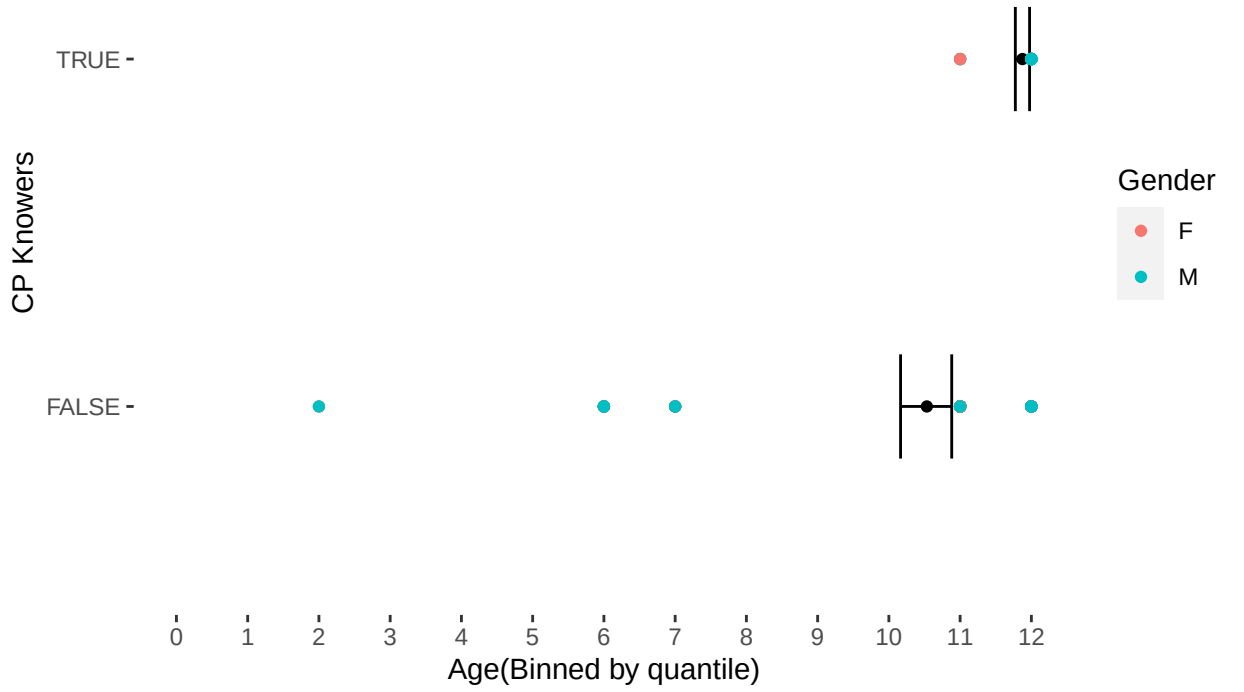
(a)

```
# Cut the part of the original dataframe with only "Tsimane" samples
d.tsimane <- subset(d, d$Language == "Tsimane")

ggplot(d.tsimane, aes(x=AgeByQuantile, y=IsCPKnowers)) +
  # geom_bar(aes(x=AgeByQuantile, y=CP_mean), stat='identity') +
  stat_summary(fun.data = mean_cl_boot, geom = "errorbar", width=.3) +
  stat_summary(fun.y = mean, geom = "point") +
  geom_point(aes(color = Gender)) +
  stat_smooth(method = "glm", method.args = list(family = binomial)) +
  xlim(0, 12) +
  scale_x_continuous(limits = c(0, 12), breaks = seq(0, 12, by = 1)) +
  xlab("Age(Binned by quantile)") +
  ylab("CP Knowers") +
  theme(panel.background = element_rect(fill = "transparent"),
        plot.background = element_rect(fill = "transparent", color = NA),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank())
```

```
## Scale for x is already present.
## Adding another scale for x, which will replace the existing scale.
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: glm.fit: algorithm did not converge
```

```
## Warning: Computation failed in `stat_smooth()`
## Caused by error:
## ! y values must be 0 <= y <= 1
```



(b) From the chart I observed that all female in "Tsimane" are the "CP Konwer". I expect to see in the logistic regression that the intercept

(c)

```
# Turn the gender predictor
d.tsimane$IsMale <- is.element(d.tsimane$Gender, "M")

# Conduct the glm
g.tsimane <- glm( CPKnower.bin ~ Age*IsMale + IsMale + Age + 1,
                  data = d.tsimane, family = "binomial")
summary(g.tsimane)
```

```
##
## Call:
## glm(formula = CPKnower.bin ~ Age * IsMale + IsMale + Age + 1,
##     family = "binomial", data = d.tsimane)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2615  -0.5818  -0.4581  -0.2359   2.4245
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -6.30083    1.64812  -3.823 0.000132 ***
```

```
## Age             0.05693    0.01899   2.999 0.002711 **
## IsMaleTRUE     -0.52202    2.16966  -0.241 0.809865
## Age:IsMaleTRUE  0.01352    0.02527   0.535 0.592748
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 196.51  on 191  degrees of freedom
## Residual deviance: 151.86  on 188  degrees of freedom
## AIC: 159.86
##
## Number of Fisher Scoring iterations: 5
```

```r
Possibility.CP.Knower.Male <- inv_logit(-6.30083 -0.52202 + (0.05693+0.01352)*Age)
Possibility.CP.Knower.Female <- inv_logit(-6.30083 + 0.05693*Age)

# Plot
plot(Age, Possibility.CP.Knower.Male, type = "l", col = "blue")
lines(Age, Possibility.CP.Knower.Female, col = "red")
legend("topright", legend = c("Male", "Female"), col = c("blue", "red"), lty = 1)
```

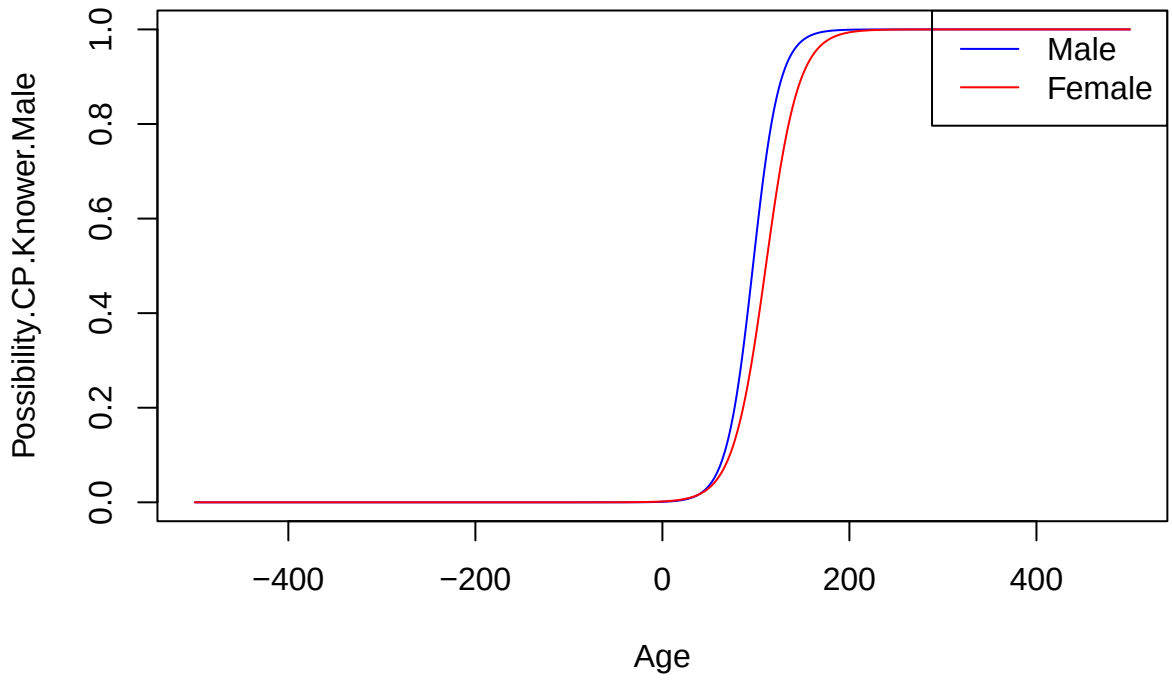


Figure 2 shows the average "Age" for "CP Knower" is longer than "non-CP Knower". Also, the difference between the min and max of "CP Knower" and "non CP Knower" is quite obvious. There's small difference in "CP Knower".

Question 7: Your collaborator keeps asking why you aren't running a linear regression, and what the relationship is between what you're doing an linear regression – after all, they use the same model formula in the glm call. Write a friendly email explaining logistic regression to them, and include argument on how it is different, and why it is more appropriate for this dataset.

Dear collaborators,

In this research, we try to figure out the relation between multi-value discrete variations(like Age) and Binary output(Like "CP-Knower or not"). Linear Regression is hard to fit that relation because it's not a direct linear relation. However, Logistic Regression is a good way to project one linear relation to a binary output. Because the output can be seen as possibility of the outcome, so it's well applied in this situation.

Best, Mingrui