# Regression

Haoyu Yue

2021/2/28

## Continuous Regression

### Import all packages and clean the environment

```
library(haven)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(pander)
library(dplyr)
rm(list=ls())
```

### Import the data

```
#
data <- read.csv("https://raw.githubusercontent.com/Group8-GovAnalyticsProject/Merging/main/final_merge
```

### State the hypotheses

```
# Hypothesis 1: Positive rate increases as median age larger.
hypo1=formula(precount_positive~median_age)
# Hypothesis 2: Positive rate increases as median age larger, income decreases.
hypo2=formula(precount_positive~median_age*median_hhold_inc)
```

**Build and compute the regression models**

```
model_1=glm(hypo1,
          data=data,
          family="gaussian")
model_2=glm(hypo2,
          data=data,
          family="gaussian")
```

**Read the result**

*Result for model/hypotheses 1*

```
pander(summary(model_1))
```

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| **(Intercept)** | 34.46 | 4.58 | 7.524 | 3.713e-13 |
| **median_age** | -0.4703 | 0.118 | -3.985 | 8.064e-05 |

(Dispersion parameter for gaussian family taken to be 159.4836 )

| | |
|---|---|
| Null deviance: | 64731 on 391 degrees of freedom |
| Residual deviance: | 62199 on 390 degrees of freedom |

*Result for model/hypotheses 2*

```
pander(summary(model_2))
```

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| **(Intercept)** | 42.1 | 11.13 | 3.782 | 0.00018 |
| **median_age** | -0.3418 | 0.2863 | -1.194 | 0.2332 |
| **median_hhold_inc** | -0.0002015 | 0.0001148 | -1.755 | 0.08011 |
| **median_age:median_hhold_inc** | 1.954e-06 | 2.833e-06 | 0.6898 | 0.4907 |

(Dispersion parameter for gaussian family taken to be 141.6419 )

| | |
|---|---|
| Null deviance: | 64731 on 391 degrees of freedom |
| Residual deviance: | 54957 on 388 degrees of freedom |

*Compare the two models and find the better one*

```
anova(model_1,model_2,test="Chisq")
```

```
## Analysis of Deviance Table
```

```
##
## Model 1: precount_positive ~ median_age
## Model 2: precount_positive ~ median_age * median_hhold_inc
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       390      62199
## 2       388      54957  2   7241.5 7.911e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Due to the p-value = 7.911e-12, these two models has statistical differences. And model 2 is better than model 1 because DF and residential are smaller.

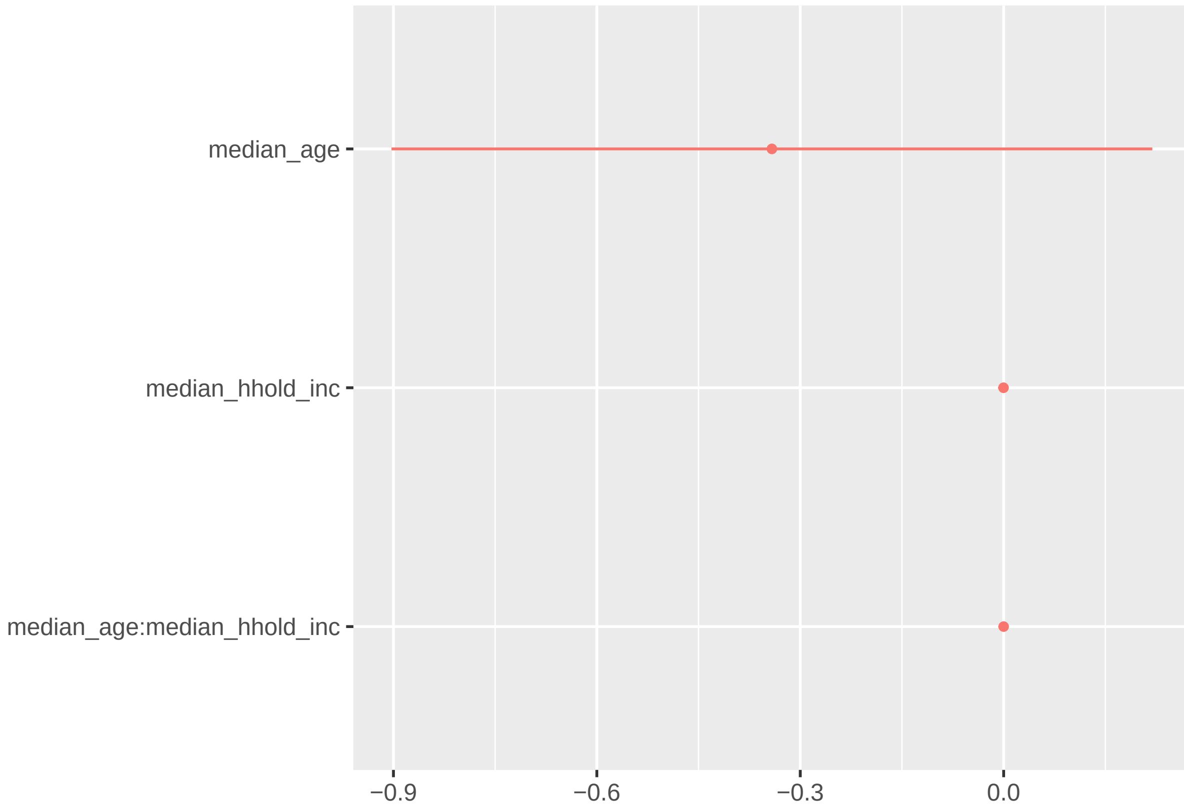
```
RSquare_1 <- rsq::rsq(model_1,adj=T)
RSquare_2 <- rsq::rsq(model_2,adj=T)
```

### Summart plots

```
library(dotwhisker)
```

```
## Registered S3 method overwritten by 'broom.mixed':
##   method      from
##   tidy.gamlss broom
```

```
library(ggplot2)
dwplot(model_2,by_2sd = F)
```




## Binary Regression

There is no binary data in our data, so we would like to add a column called high_risk, which means the percent of positive is more than the median of it.

```
data$high_risk <- ifelse(data$precount_positive>median(data$precount_positive),1,0)
data$high_risk<-factor(data$high_risk)
```

#### State the hypotheses

```
# Hypothesis 3: High risk or not has relationship with median age.
hypo3=formula(high_risk~median_age)
# Hypothesis 4: High risk or not has relationship with median age and median income.
hypo4=formula(high_risk~median_age*median_hhold_inc)
```

#### Build and compute the regression models

```
model_3=glm(hypo3,
         data=data,
         family="binomial")
```

```r
model_4=glm(hypo4,
          data=data,
          family="binomial")
```

**Read the result**

*Result for model/hypotheses 3*

```r
pander(summary(model_3))
```

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| **(Intercept)** | 3.691 | 0.8312 | 4.441 | 8.96e-06 |
| **median_age** | -0.1016 | 0.02165 | -4.692 | 2.708e-06 |

(Dispersion parameter for binomial family taken to be 1 )

| | |
|---|---|
| Null deviance: | 539.7 on 391 degrees of freedom |
| Residual deviance: | 514.7 on 390 degrees of freedom |

*Result for model/hypotheses 4*

```r
pander(summary(model_4))
```

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| **(Intercept)** | 1.772 | 2.669 | 0.6639 | 0.5067 |
| **median_age** | 0.01805 | 0.07044 | 0.2563 | 0.7978 |
| **median_hhold_inc** | 1.961e-06 | 2.884e-05 | 0.06798 | 0.9458 |
| **median_age:median_hhold_inc** | -7.555e-07 | 7.468e-07 | -1.012 | 0.3117 |

(Dispersion parameter for binomial family taken to be 1 )

| | |
|---|---|
| Null deviance: | 539.7 on 391 degrees of freedom |
| Residual deviance: | 455.8 on 388 degrees of freedom |

*Compare the two models and find the better one*

```r
lmtest::lrtest(model_3,model_4)
```

```
## Likelihood ratio test
##
## Model 1: high_risk ~ median_age
## Model 2: high_risk ~ median_age * median_hhold_inc
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   2 -257.34
## 2   4 -227.92  2 58.837  1.674e-13 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model_4 is chosen for this hypotheses.

#### Get marginal effects

```r
library(margins)
(marginINFO = margins(model_4))
```

```
## Average marginal effects

## glm(formula = hypo4, family = "binomial", data = data)

##  median_age median_hhold_inc
##    -0.01067       -5.332e-06
```

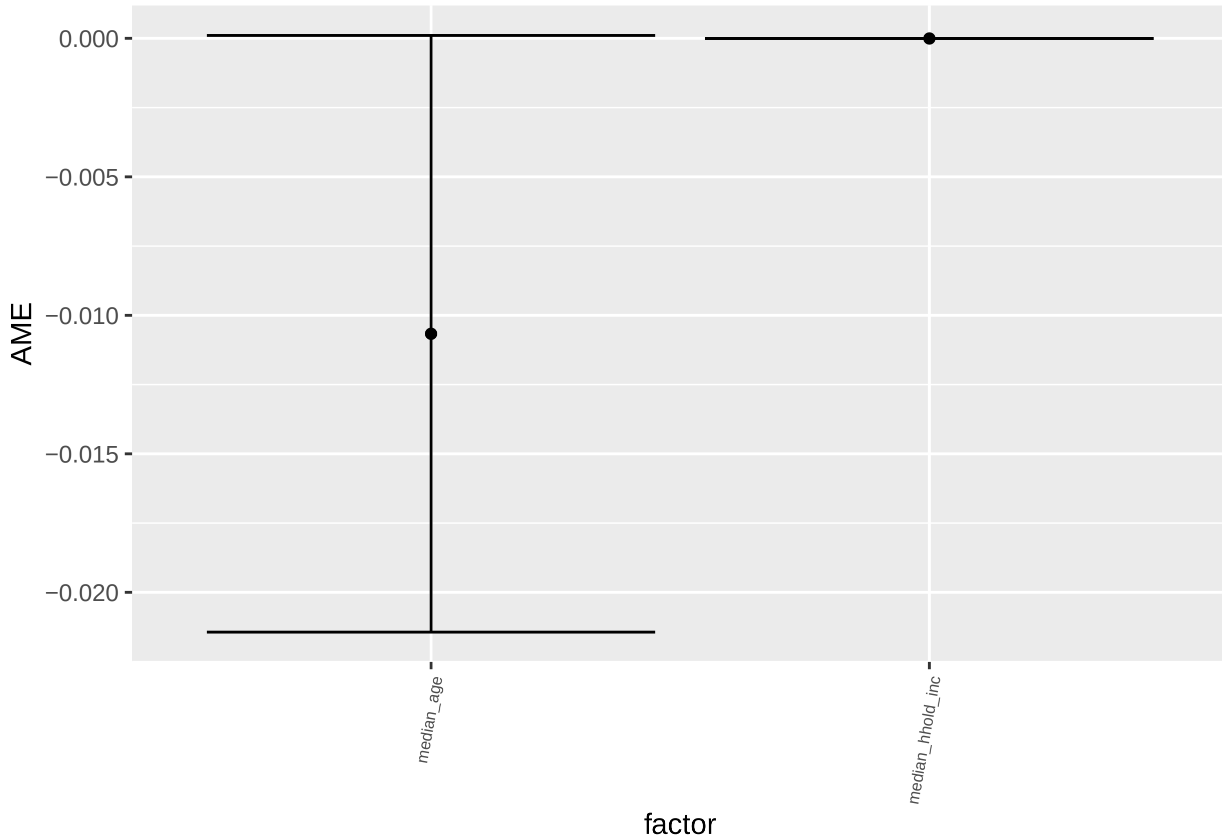
```r
(marginSUMM=summary(marginINFO))
```

```
##            factor     AME     SE       z      p   lower   upper
##        median_age -0.0107 0.0055 -1.9410 0.0523 -0.0214  0.0001
##  median_hhold_inc -0.0000 0.0000 -5.0004 0.0000 -0.0000 -0.0000
```

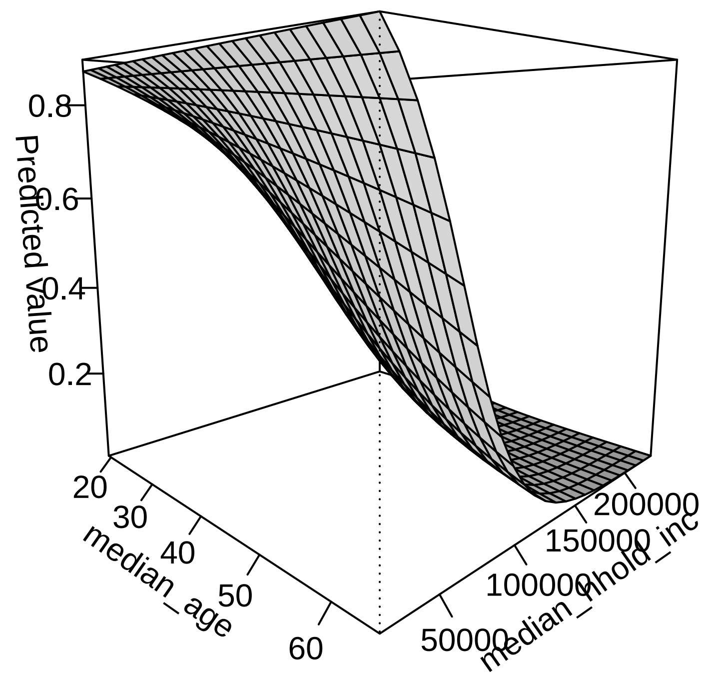
#### Get some plots

```r
base= ggplot(marginSUMM,
             aes(x=factor, y=AME))
base= base + geom_point()

plotMargins = base + theme(axis.text.x = element_text(angle = 80,size = 6,hjust = 1))
plotMargins +geom_errorbar(aes(ymin=lower, ymax=upper))
```




```r
persp(model_4)
```


0.8
0.6
0.4
0.2
Predicted Value
20
30
40
50
60
median_age
50000
100000
150000
200000
median_hhold_inc