Social_Network_Ads_logistic_regression

加载包

```
library(tidyverse)
library(effects)
library(scatterplot3d)
```

读取数据

```
social_network <- read_csv("~/workspace/Social_Network_Ads.csv")
```

```
##
## ── Column specification ────────────────────────────────────────────────────────
## cols(
##   `User ID` = col_double(),
##   Gender = col_character(),
##   Age = col_double(),
##   EstimatedSalary = col_double(),
##   Purchased = col_double()
## )
```

```
social_network
```

```
## # A tibble: 400 x 5
##     `User ID` Gender   Age EstimatedSalary Purchased
##         <dbl> <chr>  <dbl>           <dbl>     <dbl>
##  1  15624510 Male      19           19000         0
##  2  15810944 Male      35           20000         0
##  3  15668575 Female    26           43000         0
##  4  15603246 Female    27           57000         0
##  5  15804002 Male      19           76000         0
##  6  15728773 Male      27           58000         0
##  7  15598044 Female    27           84000         0
##  8  15694829 Female    32          150000         1
##  9  15600575 Male      25           33000         0
## 10  15727311 Female    35           65000         0
## # … with 390 more rows
```

检查缺失值

```
social_network %>%
    summarise_all(
        ~ sum(is.na(.))
    )
```

```
## # A tibble: 1 x 5
##   `User ID` Gender   Age EstimatedSalary Purchased
##       <int>  <int> <int>           <int>     <int>
## 1         0      0     0               0         0
```

设置训练集和测试集

```
set.seed(1234)
sample_size = round(nrow(social_network)*.70)
train <- sample_n(social_network, sample_size)
train
```

```
## # A tibble: 280 x 5
##    `User ID` Gender   Age EstimatedSalary Purchased
##        <dbl> <chr>  <dbl>           <dbl>     <dbl>
##  1  15663249 Female    52           21000         1
##  2  15601550 Female    36           54000         0
##  3  15766289 Male      27           88000         0
##  4  15665416 Female    39           71000         0
##  5  15748589 Female    45           45000         1
##  6  15725660 Male      30           87000         0
##  7  15627220 Male      39           71000         0
##  8  15582492 Male      28          123000         1
##  9  15584545 Female    32           86000         0
## 10  15657163 Male      35           58000         0
## # … with 270 more rows
```

```
sample_id <- as.numeric(rownames(train))
test <- social_network[-sample_id,]
test
```

```
## # A tibble: 120 x 5
##    `User ID` Gender   Age EstimatedSalary Purchased
##        <dbl> <chr>  <dbl>           <dbl>     <dbl>
##  1  15609669 Female    59           88000         1
##  2  15685536 Male      35           61000         0
##  3  15750447 Male      37           70000         1
##  4  15663249 Female    52           21000         1
##  5  15638646 Male      48          141000         0
##  6  15734161 Female    37           93000         1
##  7  15631070 Female    37           62000         0
##  8  15761950 Female    48          138000         1
##  9  15649668 Male      41           79000         0
## 10  15713912 Female    37           78000         1
## # … with 110 more rows
```

将性别和购买与否设置为因子

```
train <- train %>%
  mutate(Gender = factor(Gender),
         Purchased = factor(Purchased))
train
```

```
## # A tibble: 280 x 5
##    `User ID` Gender   Age EstimatedSalary Purchased
##        <dbl> <fct>  <dbl>           <dbl> <fct>
##  1  15663249 Female    52           21000 1
##  2  15601550 Female    36           54000 0
##  3  15766289 Male      27           88000 0
##  4  15665416 Female    39           71000 0
##  5  15748589 Female    45           45000 1
##  6  15725660 Male      30           87000 0
##  7  15627220 Male      39           71000 0
##  8  15582492 Male      28          123000 1
##  9  15584545 Female    32           86000 0
## 10  15657163 Male      35           58000 0
## # … with 270 more rows
```

#计算购买的概率和方差

```
train %>%
  count(Purchased)
```

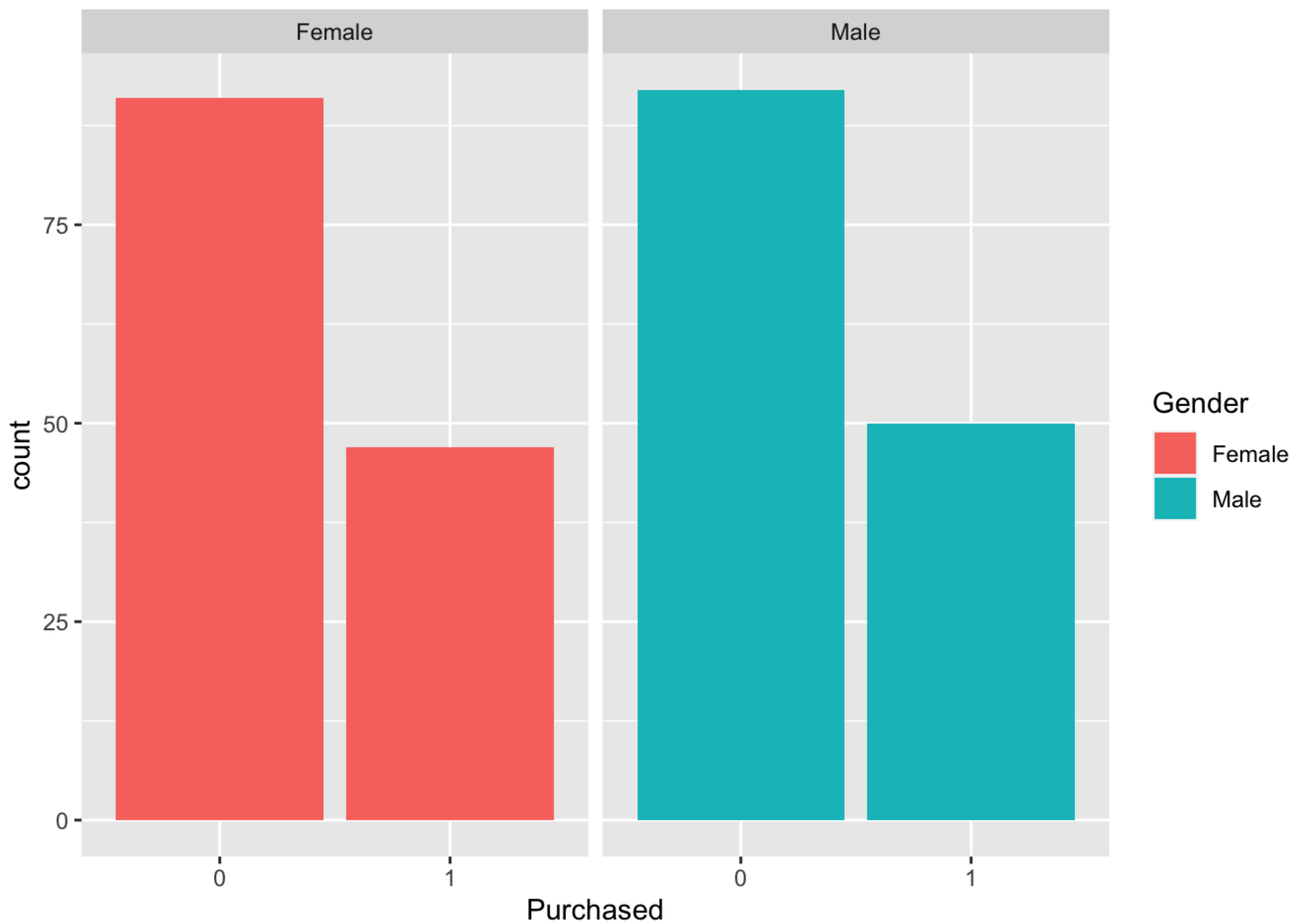
```
## # A tibble: 2 x 2
##   Purchased     n
##   <fct>     <int>
## 1 0           183
## 2 1            97
```

```
prob <- tibble(p = 143/280,
       q = 1-p,
       var = 280*p*q)
prob
```

```
## # A tibble: 1 x 3
##       p     q   var
##   <dbl> <dbl> <dbl>
## 1 0.511 0.489  70.0
```

按性别对购买行为分组，查看购买差异分布

```
train %>%
ggplot(aes(Purchased,  fill = Gender))+
  geom_bar()+
  facet_grid(.~Gender)
```



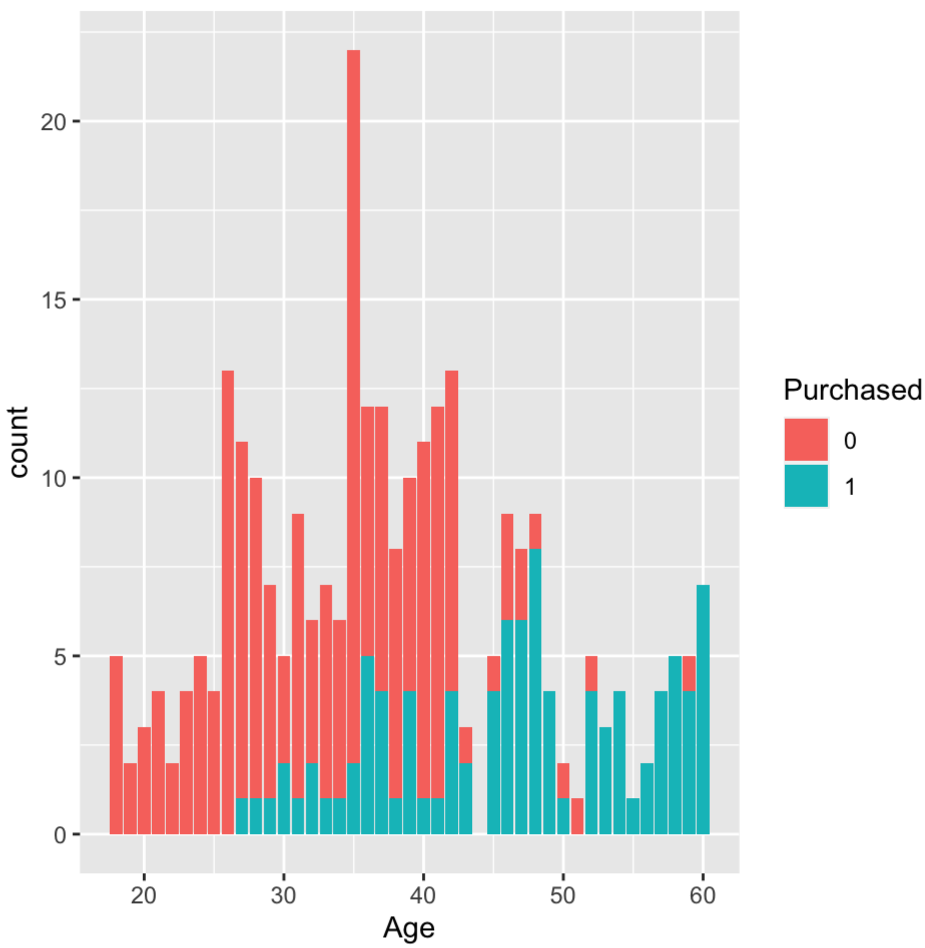
性别对购买与否影响不大。

按年龄对购买行为进行分组，查看差异

```
train %>%
  count(Age)
```

```
## # A tibble: 42 x 2
##      Age     n
##    <dbl> <int>
##  1    18     5
##  2    19     2
##  3    20     3
##  4    21     4
##  5    22     2
##  6    23     4
##  7    24     5
##  8    25     4
##  9    26    13
## 10    27    11
## # … with 32 more rows
```

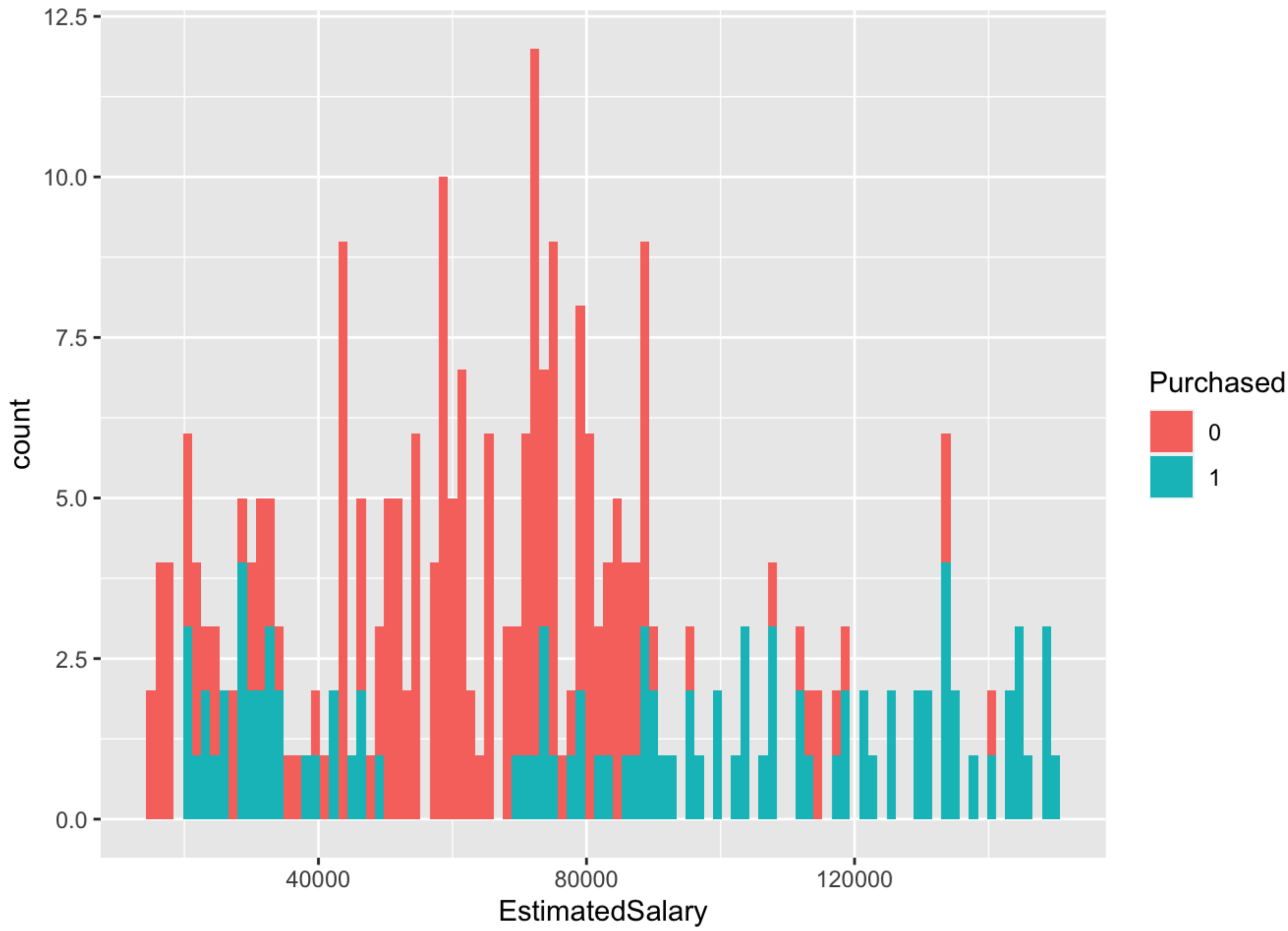
```
 train %>%
  ggplot(aes(Age, fill = Purchased))+
  geom_bar()
```



26岁一下无人购买，26-42岁不购买人数多于购买人数，43岁以上为主要购买人数

按估计薪水对购买行为进行分组，查看分布差异

```
train %>%
  ggplot(aes(EstimatedSalary, fill = Purchased))+
  geom_histogram( bins = 100)
```



薪水2万-5万 和7万-8.5万，非购买人数多于购买人数，8.5万以上购买人数多于非购买人数，但有几处情况不是这样

薪水2万以下和5万-6.5万无人购买

```
train %>%
  ggplot(aes(Age, EstimatedSalary, color = Purchased)) +
  geom_point()
```

大致上薪水8.2万一下且年龄小于41岁无人购买。

逻辑回归模型

模型1

```
mod1 <- glm(Purchased~EstimatedSalary+Age+Gender, family = binomial(link = "logit"), dat
a = train)
summary(mod1)
```

```
##
## Call:
## glm(formula = Purchased ~ EstimatedSalary + Age + Gender, family = binomial(link = "l
ogit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7332  -0.5389  -0.1898   0.3825   2.4572
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -1.207e+01  1.501e+00  -8.044 8.72e-16 ***
## EstimatedSalary  3.243e-05  6.213e-06   5.219 1.79e-07 ***
## Age              2.219e-01  2.923e-02   7.593 3.12e-14 ***
## GenderMale       5.138e-01  3.614e-01   1.422    0.155
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 361.32  on 279  degrees of freedom
## Residual deviance: 198.98  on 276  degrees of freedom
## AIC: 206.98
##
## Number of Fisher Scoring iterations: 6
```

性别的p值过大，也验证了图形中反映的情况

模型2

```
mod2 <- glm(Purchased~EstimatedSalary+Age, family = binomial(link = "logit"), data = tra
in)
summary(mod2)
```

```
##
## Call:
## glm(formula = Purchased ~ EstimatedSalary + Age, family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7962  -0.5737  -0.2025   0.3915   2.3285
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -1.158e+01  1.431e+00  -8.094 5.79e-16 ***
## EstimatedSalary  3.207e-05  6.190e-06   5.181 2.20e-07 ***
## Age              2.171e-01  2.870e-02   7.567 3.83e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 361.32  on 279  degrees of freedom
## Residual deviance: 201.04  on 277  degrees of freedom
## AIC: 207.04
##
## Number of Fisher Scoring iterations: 6
```

模型2的p值均小于0.1%，且AIC值并没有降低。

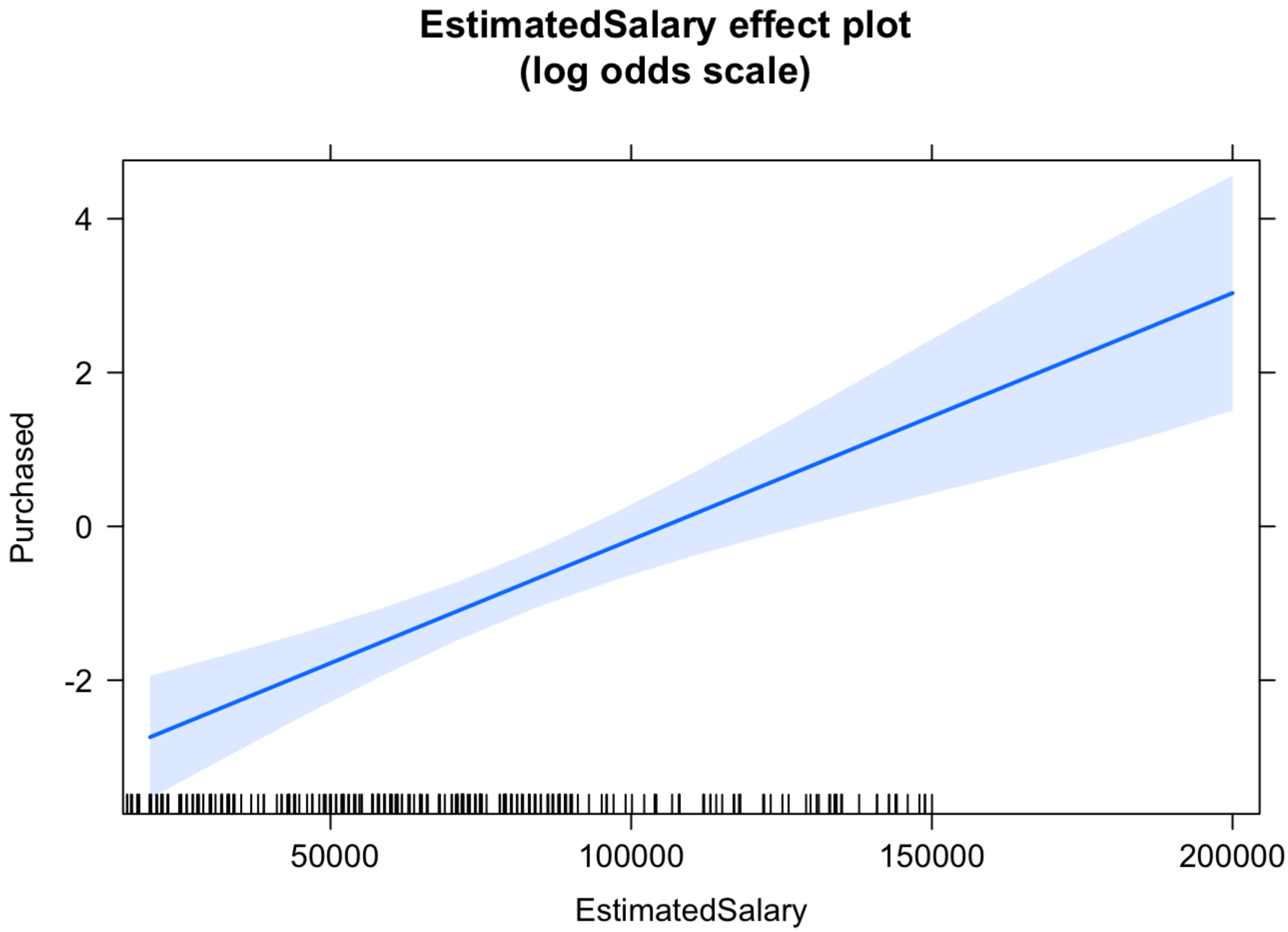
指数化模型参数

```
coef(mod2) %>%
  exp() %>%
  round(digits = 6)
```

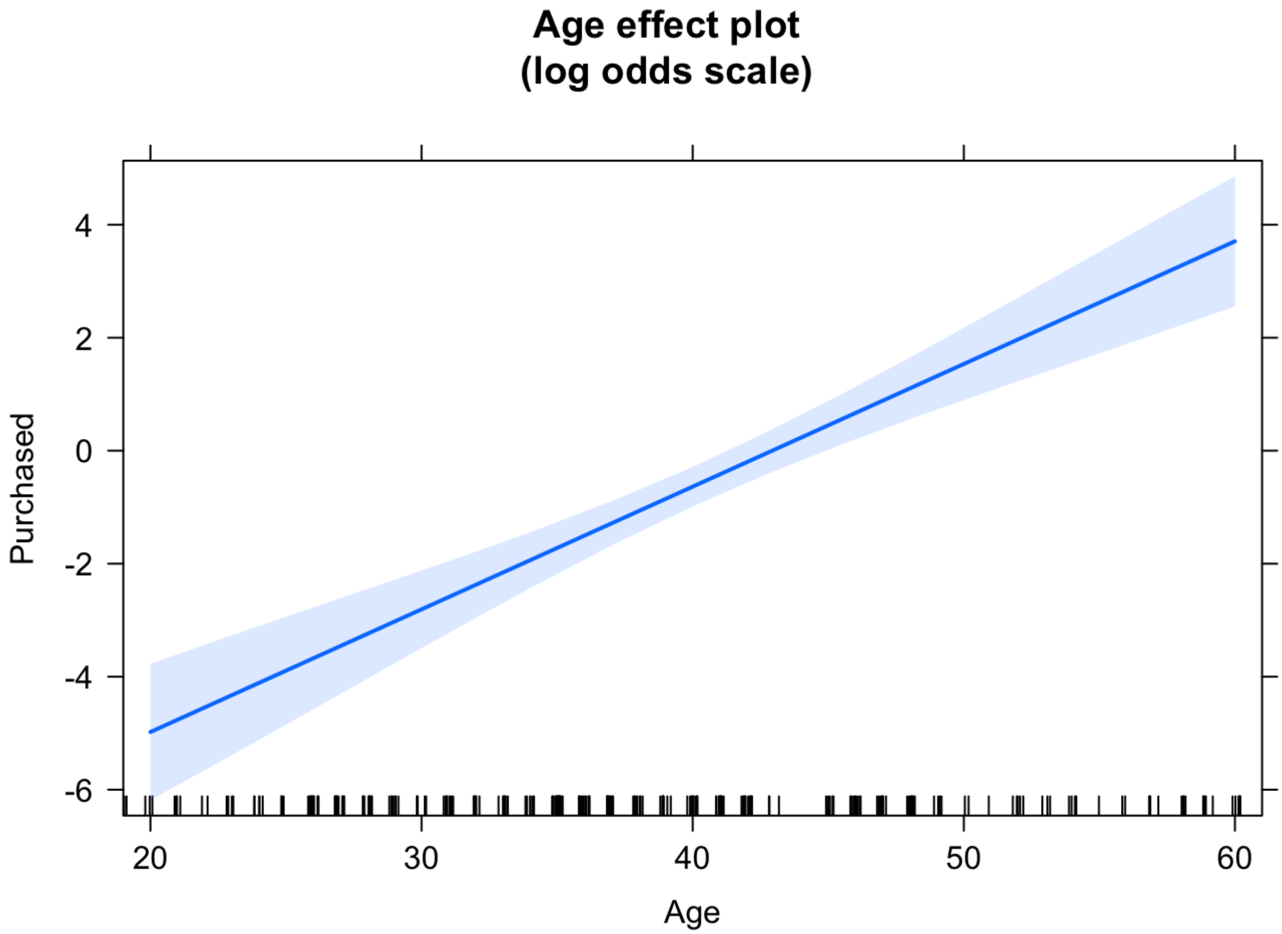
```
##     (Intercept) EstimatedSalary             Age
##        0.000009        1.000032        1.242508
```

在控制薪水不变的情况下，年龄每增加一个单位，购买的概率增加1的1.24次方。

```
effect_link <- Effect("EstimatedSalary", mod = mod2)
plot(effect_link,
  type = "link",
  main = "EstimatedSalary effect plot\n(log odds scale)"
)
```




```
effect_link <- Effect("Age", mod = mod2)
plot(effect_link,
  type = "link",
  main = "Age effect plot\n(log odds scale)"
)
```

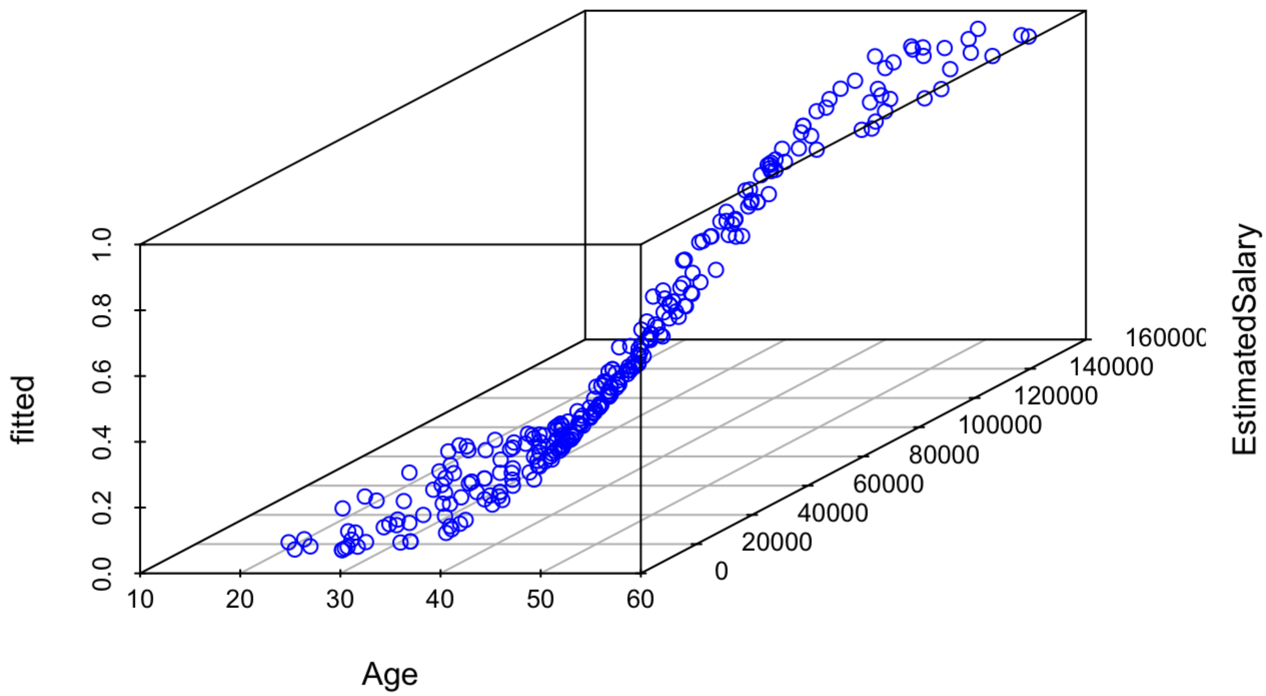


EstimatedSalary >120000 hava more residuals, and Age<30 or Age>50 have more residuals too.

```
train %>%
  mutate(fitted = fitted(mod2))
```

```
## # A tibble: 280 x 6
##    `User ID` Gender   Age EstimatedSalary Purchased fitted
##        <dbl> <fct>  <dbl>           <dbl> <fct>      <dbl>
##  1  15663249 Female    52           21000 1         0.594
##  2  15601550 Female    36           54000 0         0.116
##  3  15766289 Male      27           88000 0         0.0522
##  4  15665416 Female    39           71000 0         0.302
##  5  15748589 Female    45           45000 1         0.409
##  6  15725660 Male      30           87000 0         0.0928
##  7  15627220 Male      39           71000 0         0.302
##  8  15582492 Male      28          123000 1         0.174
##  9  15584545 Female    32           86000 0         0.133
## 10  15657163 Male      35           58000 0         0.107
## # … with 270 more rows
```

```
train %>%
  mutate(fitted = fitted(mod2)) %>%
  select(Age, EstimatedSalary, fitted) %>%
  scatterplot3d(color = "blue")
```



age<50 and EstimatedSalary < 70000 have the probility that less than 0.8

```
prob<-predict(mod2,test,type="response")
test
```

```
## # A tibble: 120 x 5
##    `User ID` Gender   Age EstimatedSalary Purchased
##        <dbl> <chr>  <dbl>           <dbl>     <dbl>
##  1  15609669 Female    59           88000         1
##  2  15685536 Male      35           61000         0
##  3  15750447 Male      37           70000         1
##  4  15663249 Female    52           21000         1
##  5  15638646 Male      48          141000         0
##  6  15734161 Female    37           93000         1
##  7  15631070 Female    37           62000         0
##  8  15761950 Female    48          138000         1
##  9  15649668 Male      41           79000         0
## 10  15713912 Female    37           78000         1
## # … with 110 more rows
```