Problem Set 2

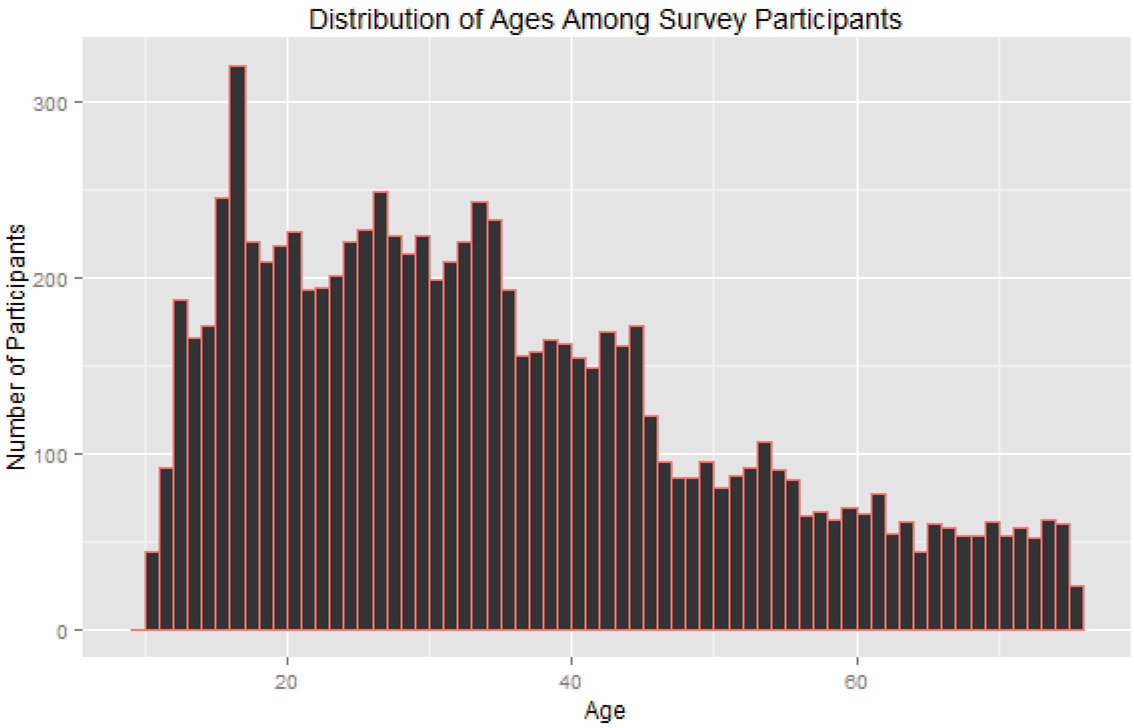
Working in collaboration with: Christine Malinowski, Kathryn Cuff, Nic Dobbins
Estimated time: 18 hours

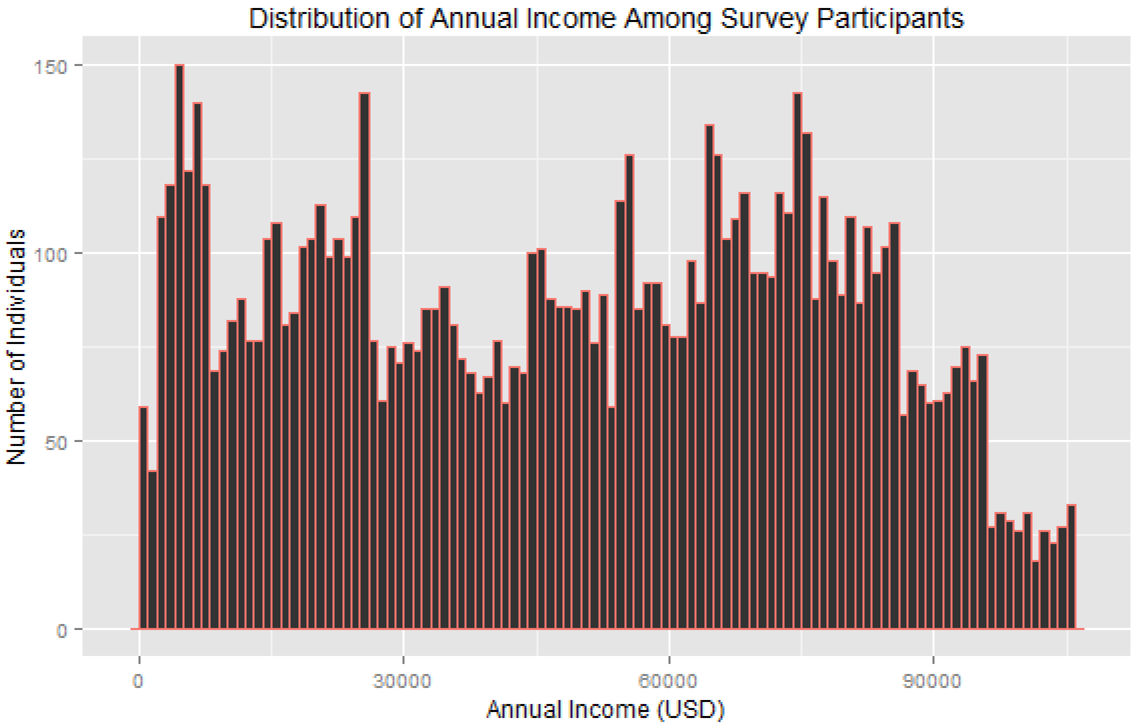
1. Summary Statistics

1.A Summary Statistics for Marketing Dataset

Variable	Median	Mean	SD	NA
Sex	2 (Female)	1.547	.4978261	0
Marital Status	3 (Divorced or separated)	3.031	1.809343	160
Education	4 (1 to 3 years of college)	3.835	1.242391	86
Occupation	4 (Clerical/Service Worker)	3.788	2.544717	136
Tenure in Bay Area	5 (More than ten years)	4.198	1.226254	913
Dual Incomes (if married)	1 (Not Married)	1.545	.7395653	0
Household Size	3	2.852	1.535914	375
Under 18 in Household	0	0.6669	1.083886	0
Householder Status	2 (Rent)	1.837	.7443959	240
Type of Home	1 (House)	1.856	1.145279	357
Ethnicity	7 (White)	5.956	1.757631	68
Language in home?	1 (English)	1.127	.4144111	359
Annual Household Income	50,599	49,118	28677.74	0
Age	32	34.43	16.22612	0

1.B Histograms





2. The T-Test

2.A Compare income of men and women

The average income of men is $50138.47 vs. $48272.20 for women. The p-value of 0.002082 tells us that there is a 99.7918% probability that these income averages are not different by random chance. This falls well above the generally accepted probability of 95% to be highly confident.

2.B Compare income of people who have lived in Bay Area 4-6 years vs. 7-10 years

The average income of people living in the Bay Area 4-6 years is $48014.53 vs. $50005.76 for those who have lived in the Bay Area 7-10 years. The p-value of 0.1742 tells us that there is a 82.58% probability that these income averages are not different at this rate by random chance. This falls well below the generally accepted probability of 95%. Said another way, there is a 17.42% possibility that the differences in income are due to random chance. This is generally a low degree of confidence.

2.C Compare the number of household members under the age of 18 for men and women

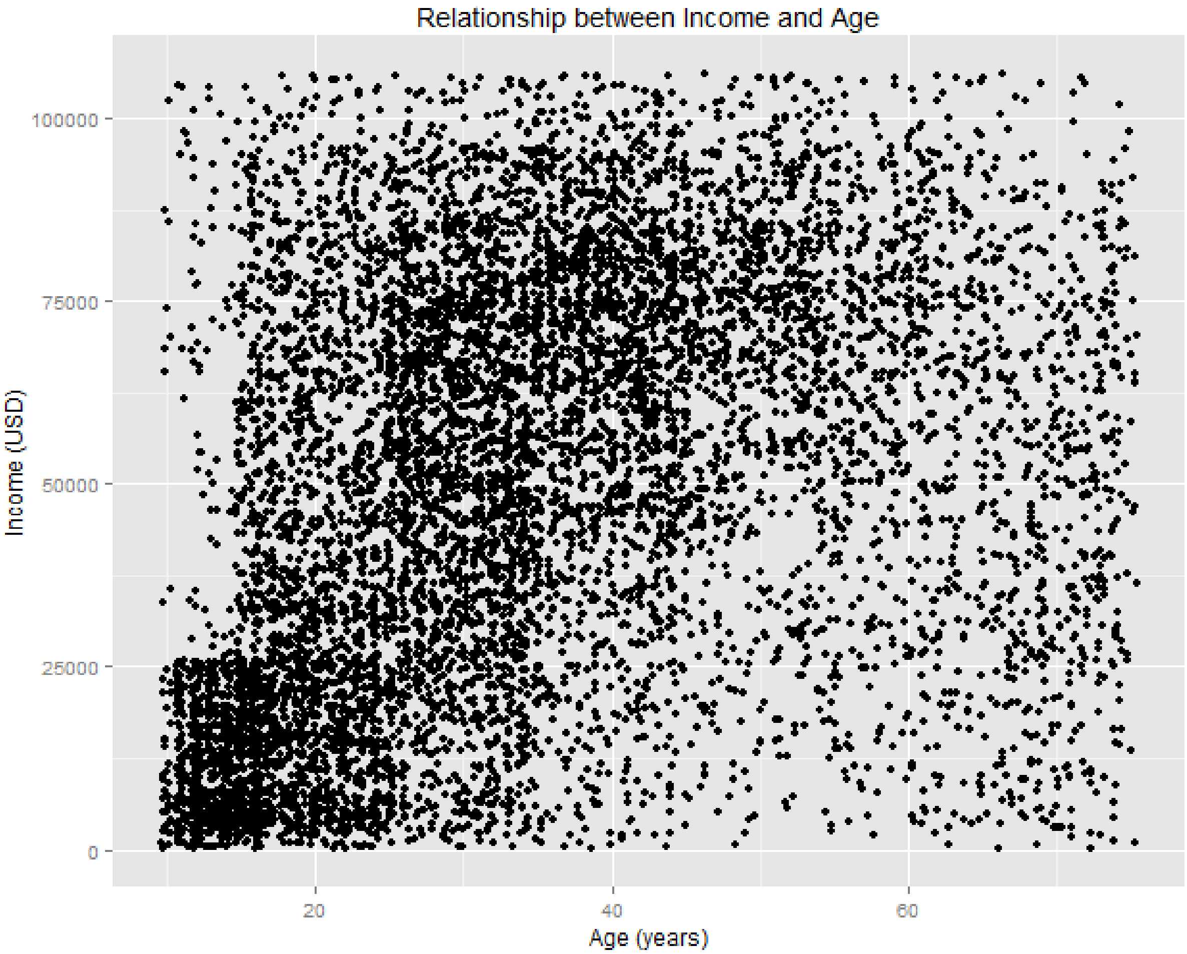
The average number of household members under the age of 18 in the households of men is 0.5796319 vs. 0.7391216 in the households of women. The p-value of 3.017e-12 tells us that there is well over a 99.9999% probability that these averages are not different by random chance. This falls well above the generally accepted probability of 95% to be highly confident.

2.D Compare income of people with and without children

The average income of households with children is $46366.03 vs. $50689.45 for households without children. The p-value of 2.198e-11 tells us that there is well over a 99.9999% probability that these averages are not different by random chance. This falls well above the generally accepted probability of 95% to be highly confident.

3. Bivariate Regression

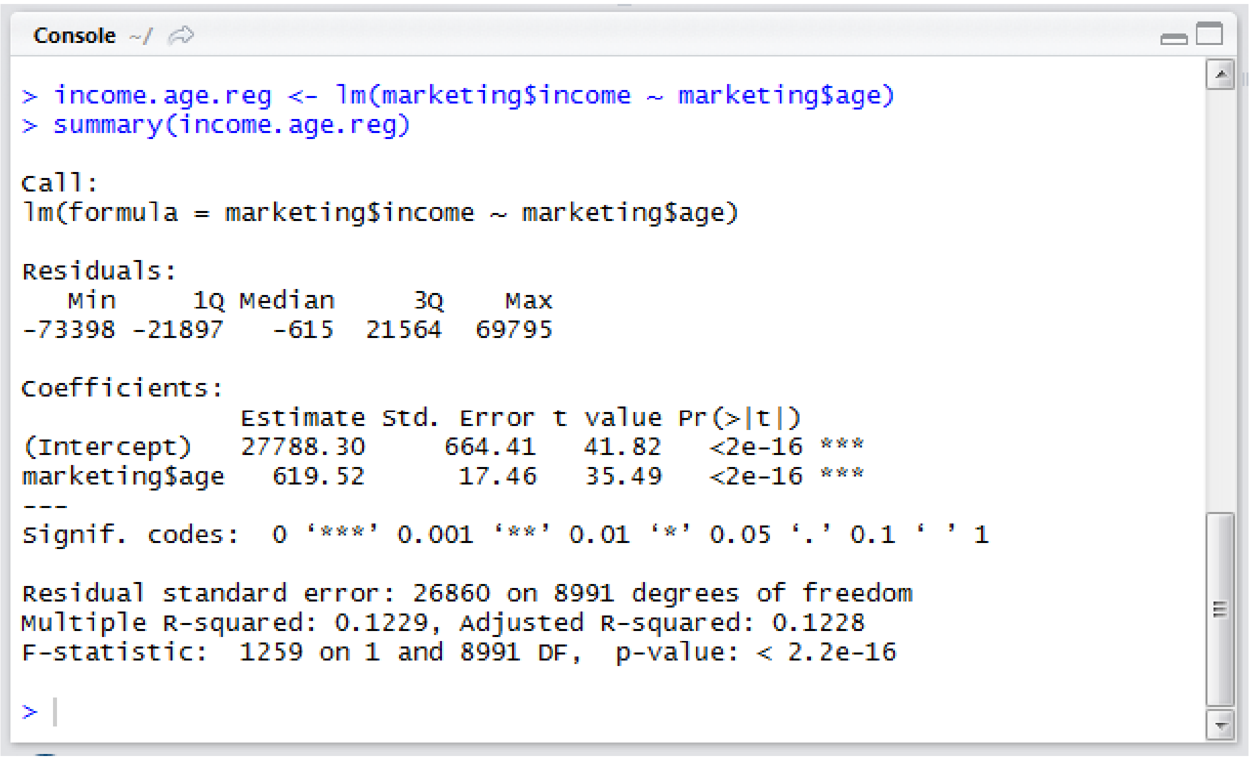
3.A Plot the relationship between income and age



3.B Estimate the regression model

$income_i = \alpha + \beta * age_i + error_i$

$income_i = 27788.30 + 619.52 * age_i + error_i$

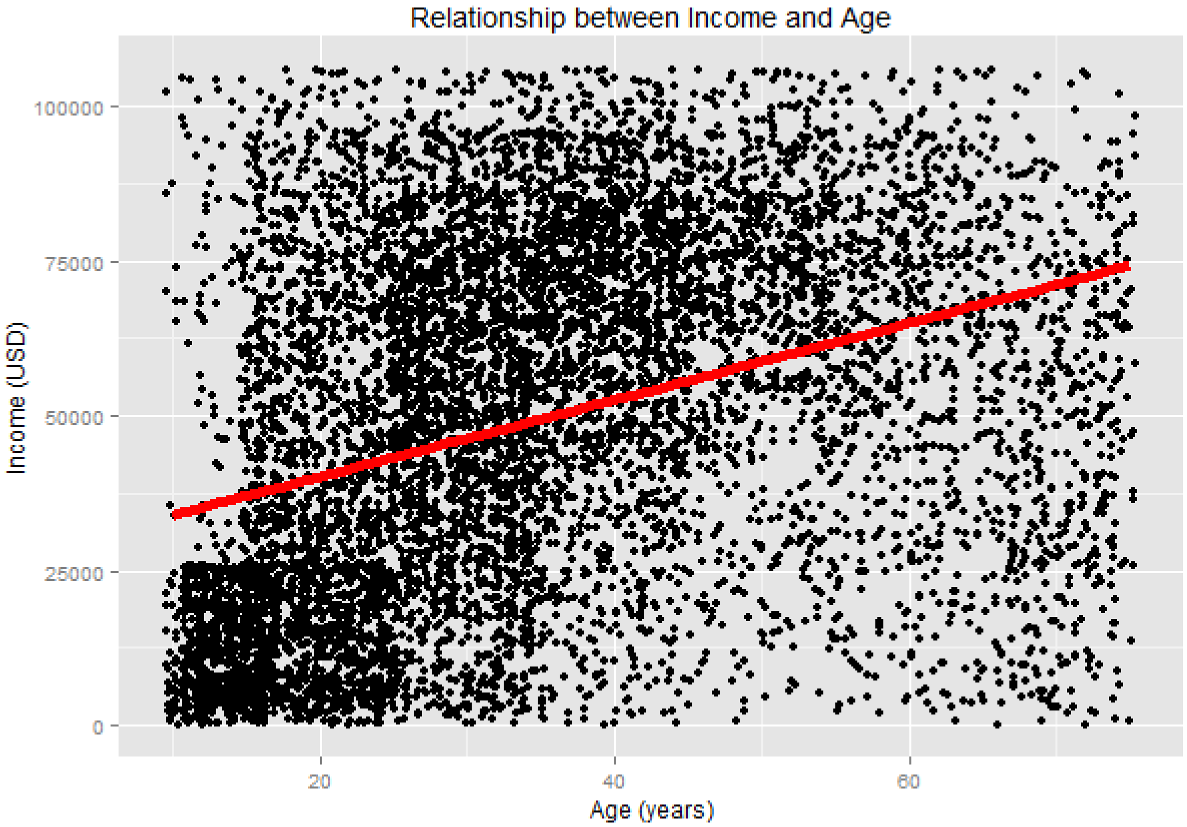


```
> income.age.reg <- lm(marketing$income ~ marketing$age)
> summary(income.age.reg)

Call:
lm(formula = marketing$income ~ marketing$age)

Residuals:
   Min     1Q Median     3Q    Max 
-73398 -21897   -615  21564  69795 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   27788.30     664.41   41.82   <2e-16 ***
marketing$age   619.52      17.46   35.49   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 26860 on 8991 degrees of freedom
Multiple R-squared: 0.1229, Adjusted R-squared: 0.1228 
F-statistic:  1259 on 1 and 8991 DF,  p-value: < 2.2e-16 

> 
```

3.C Explain the relationship between age and income

By simply looking at the plot, one can see a high density of points starting in the lower left corner. This tells us that individuals below the age of 30 tend to make less money. As people get older, say up to age 50, they tend to make more money, as indicated by the density of points in the upper mid area of the graph. From age 50 and older, income becomes more disparate. Because the p-value of income by age is very small (less than 0.000001) there is a high degree of confidence that income increases with age.

3.D Add a regression line

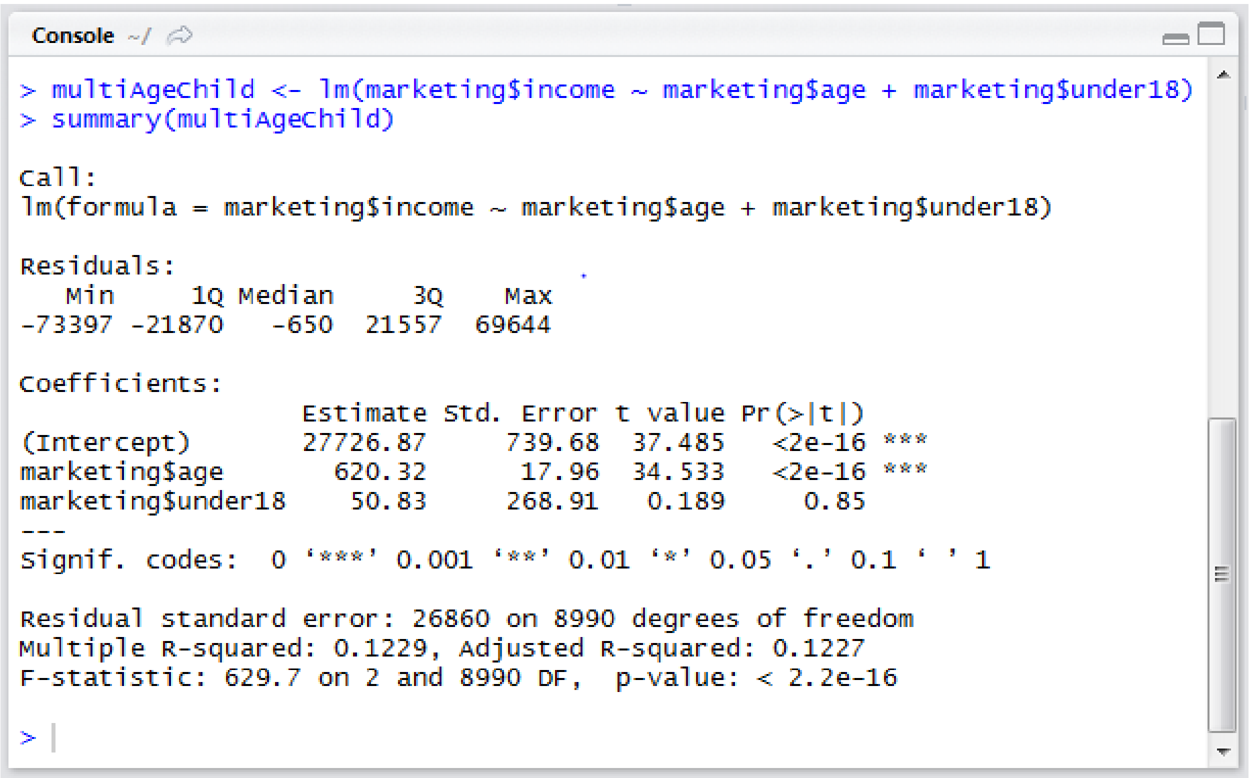


3.E Re-estimate using the quadratic model

4. Multivariate Regression

4.A Incorporate number of children in the household into the regression model

$income_i = \alpha + \beta * age_i + \gamma * children_i + error_i$
$income_i = 27726.87 + 620.32 * age_i + 50.83 * children_i + error_i$



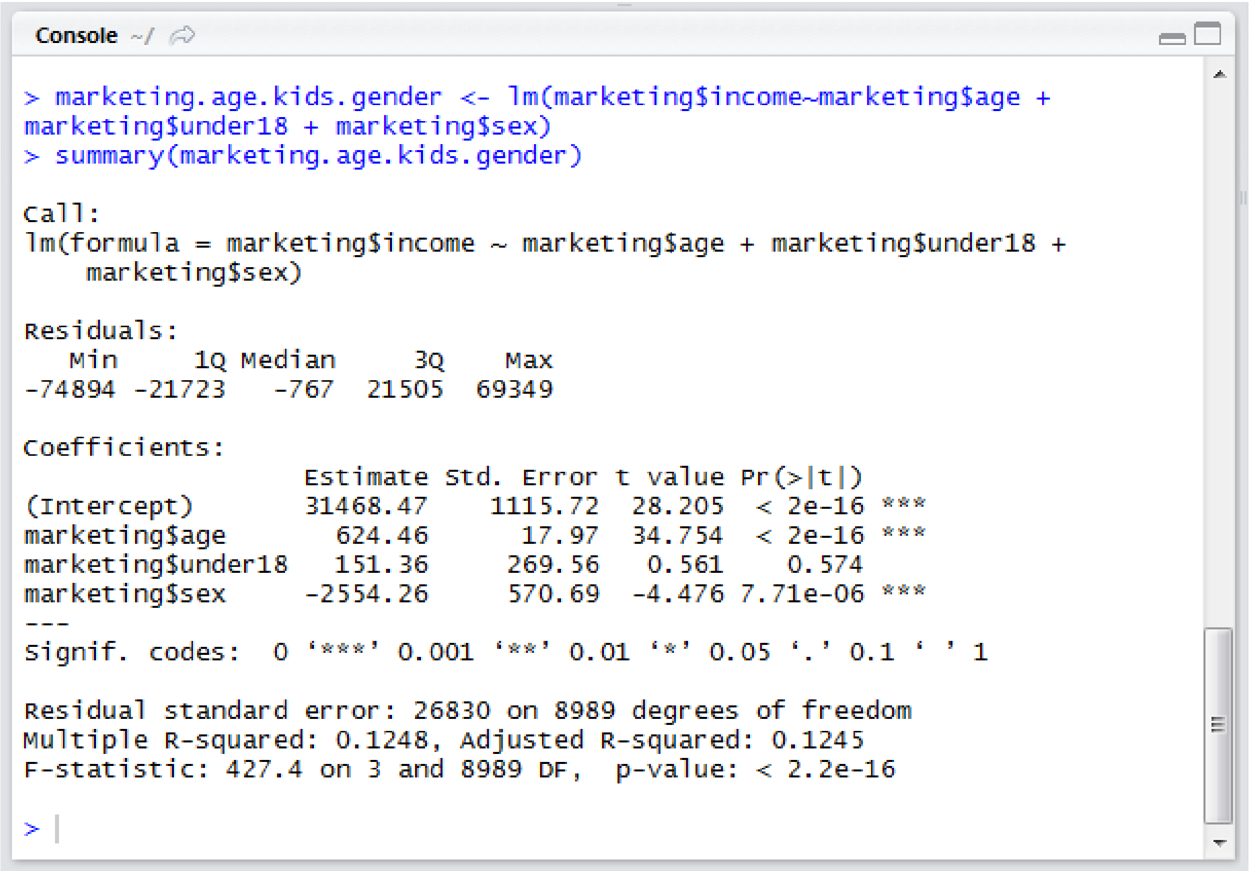
```
> multiAgeChild <- lm(marketing$income ~ marketing$age + marketing$under18)
> summary(multiAgeChild)

Call:
lm(formula = marketing$income ~ marketing$age + marketing$under18)

Residuals:
   Min     1Q Median     3Q    Max 
-73397 -21870   -650  21557  69644 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       27726.87     739.68  37.485   <2e-16 ***
marketing$age       620.32      17.96  34.533   <2e-16 ***
marketing$under18    50.83     268.91   0.189     0.85    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26860 on 8990 degrees of freedom
Multiple R-squared: 0.1229, Adjusted R-squared: 0.1227 
F-statistic: 629.7 on 2 and 8990 DF,  p-value: < 2.2e-16

> 
```

Compared to running the formula simply for age, we can see that each child increases income by an estimated $50.83. Age has a similar influence as the previous formula, with estimated income now at 620.32/year versus 619.52/year previously. The R squared is the same at 0.1229 so adding children as a variable did not make this model a better fit than the previous model.

4.B Incorporate gender

$income_i = \alpha + \beta * age_i + \gamma * children_i + \delta * sex_i + error_i$

$income_i = 31468.47 + 624.46 * age_i + 151.36 * children_i - 2554.26 * sex_i + error_i$



```
> marketing.age.kids.gender <- lm(marketing$income~marketing$age +
marketing$under18 + marketing$sex)
> summary(marketing.age.kids.gender)

Call:
lm(formula = marketing$income ~ marketing$age + marketing$under18 +
    marketing$sex)

Residuals:
   Min     1Q Median     3Q    Max
-74894 -21723   -767  21505  69349

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       31468.47    1115.72  28.205  < 2e-16 ***
marketing$age       624.46      17.97  34.754  < 2e-16 ***
marketing$under18   151.36     269.56   0.561    0.574
marketing$sex     -2554.26     570.69  -4.476 7.71e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26830 on 8989 degrees of freedom
Multiple R-squared: 0.1248, Adjusted R-squared: 0.1245
F-statistic: 427.4 on 3 and 8989 DF,  p-value: < 2.2e-16

> 
```

4.C Answer the following

What is the expected income of a 52-year old male with 3 kids?
$61840.21

What is the expected income of a 22-year old female with 1 kids?
$40249.43

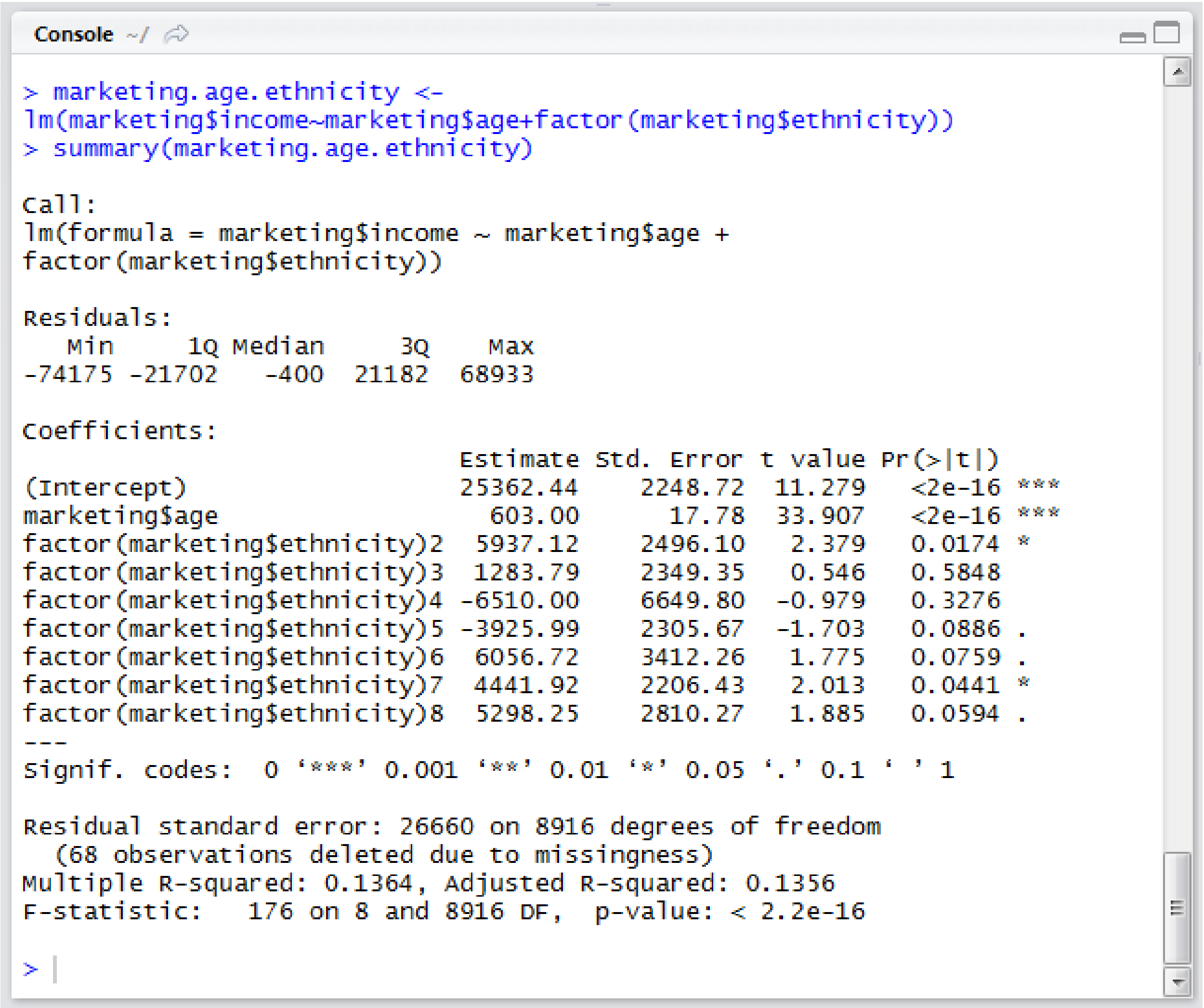
What is the expected income of a 23-year old female with 0 kids?
$40722.53

What is the difference in income between a 40-year old male with 2 kids and a 40-year old female with 2 kids?
$2554.26

4.D How age and ethnicity impacts income

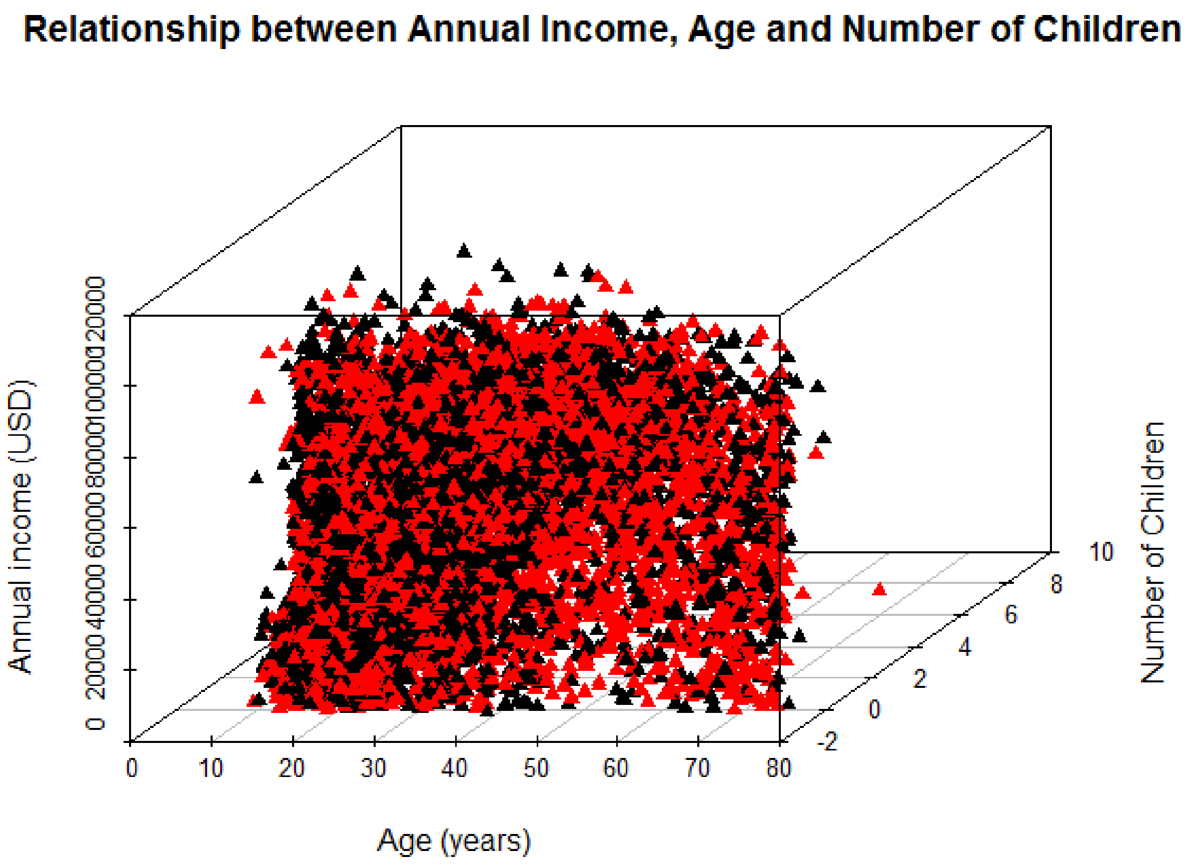
$income_i = \alpha + \beta * age_i + \gamma * ethnicity + error_i$

$income_i = 3127.39 + 412.23 * age_i + 8299.43 * ethnicity + error_i$



```
> marketing.age.ethnicity <-
lm(marketing$income~marketing$age+factor(marketing$ethnicity))
> summary(marketing.age.ethnicity)

Call:
lm(formula = marketing$income ~ marketing$age + 
factor(marketing$ethnicity))

Residuals:
   Min     1Q Median     3Q    Max 
-74175 -21702   -400  21182  68933 

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                  25362.44    2248.72  11.279   <2e-16 ***
marketing$age                  603.00      17.78  33.907   <2e-16 ***
factor(marketing$ethnicity)2  5937.12    2496.10   2.379   0.0174 *  
factor(marketing$ethnicity)3  1283.79    2349.35   0.546   0.5848    
factor(marketing$ethnicity)4 -6510.00    6649.80  -0.979   0.3276    
factor(marketing$ethnicity)5 -3925.99    2305.67  -1.703   0.0886 .  
factor(marketing$ethnicity)6  6056.72    3412.26   1.775   0.0759 .  
factor(marketing$ethnicity)7  4441.92    2206.43   2.013   0.0441 *  
factor(marketing$ethnicity)8  5298.25    2810.27   1.885   0.0594 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26660 on 8916 degrees of freedom
  (68 observations deleted due to missingness)
Multiple R-squared: 0.1364, Adjusted R-squared: 0.1356 
F-statistic:   176 on 8 and 8916 DF,  p-value: < 2.2e-16

> 
```

4.E 3D Plot



Relationship between Annual Income, Age and Number of Children

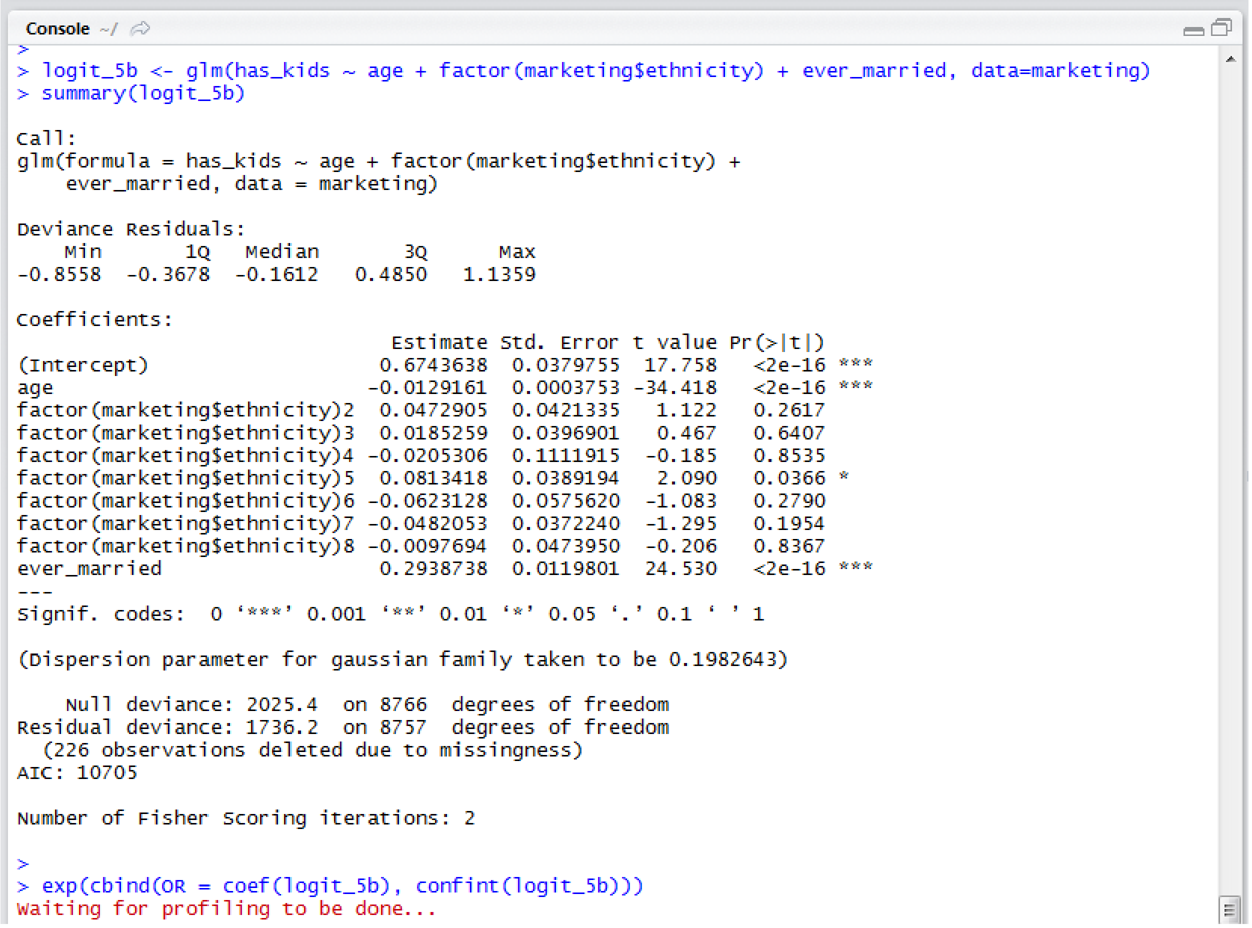
5. Logistic Regression

5.A Estimate the regression model

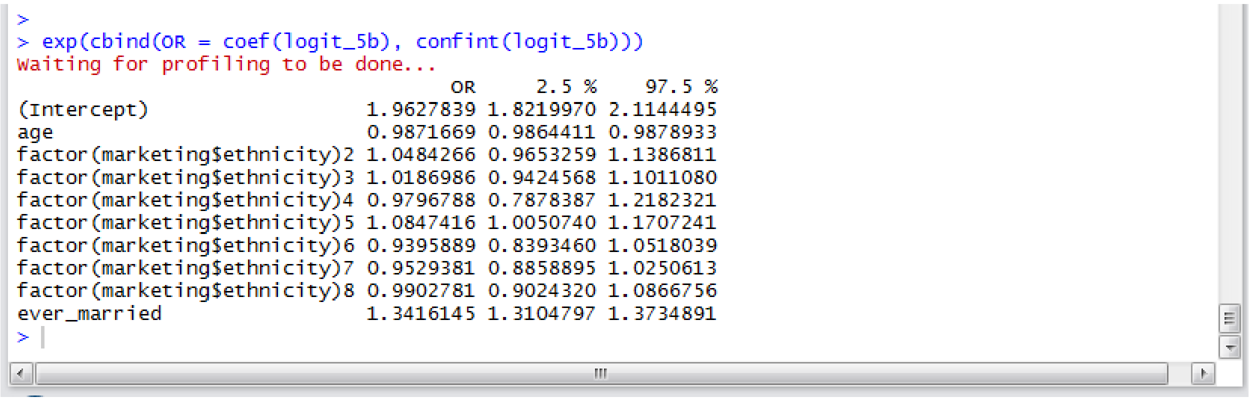
$has\_kids_i = \alpha + \beta * age_i + \delta * ever\_married_i + error_i$.
$has\_kids_i = 0.6753122 – 0.0134815 * age_i + 0.2971912 * ever\_married_i + 0.0113049$.

For every year of age, the probability of having kids goes up by 0.986609. The probability of having kids once married, whether currently married or not, go up by 1.346073 for every year of age.

5.B Improve for efficiency



```
>
> logit_5b <- glm(has_kids ~ age + factor(marketing$ethnicity) + ever_married, data=marketing)
> summary(logit_5b)

Call:
glm(formula = has_kids ~ age + factor(marketing$ethnicity) +
    ever_married, data = marketing)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8558  -0.3678  -0.1612   0.4850   1.1359

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  0.6743638  0.0379755  17.758   <2e-16 ***
age                         -0.0129161  0.0003753 -34.418   <2e-16 ***
factor(marketing$ethnicity)2  0.0472905  0.0421335   1.122   0.2617
factor(marketing$ethnicity)3  0.0185259  0.0396901   0.467   0.6407
factor(marketing$ethnicity)4 -0.0205306  0.1111915  -0.185   0.8535
factor(marketing$ethnicity)5  0.0813418  0.0389194   2.090   0.0366 *
factor(marketing$ethnicity)6 -0.0623128  0.0575620  -1.083   0.2790
factor(marketing$ethnicity)7 -0.0482053  0.0372240  -1.295   0.1954
factor(marketing$ethnicity)8 -0.0097694  0.0473950  -0.206   0.8367
ever_married                 0.2938738  0.0119801  24.530   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.1982643)

    Null deviance: 2025.4  on 8766  degrees of freedom
Residual deviance: 1736.2  on 8757  degrees of freedom
  (226 observations deleted due to missingness)
AIC: 10705

Number of Fisher Scoring iterations: 2

>
> exp(cbind(OR = coef(logit_5b), confint(logit_5b)))
Waiting for profiling to be done...
```



```
>
> exp(cbind(OR = coef(logit_5b), confint(logit_5b)))
Waiting for profiling to be done...
                                    OR     2.5 %    97.5 %
(Intercept)                  1.9627839 1.8219970 2.1144495
age                          0.9871669 0.9864411 0.9878933
factor(marketing$ethnicity)2 1.0484266 0.9653259 1.1386811
factor(marketing$ethnicity)3 1.0186986 0.9424568 1.1011080
factor(marketing$ethnicity)4 0.9796788 0.7878387 1.2182321
factor(marketing$ethnicity)5 1.0847416 1.0050740 1.1707241
factor(marketing$ethnicity)6 0.9395889 0.8393460 1.0518039
factor(marketing$ethnicity)7 0.9529381 0.8858895 1.0250613
factor(marketing$ethnicity)8 0.9902781 0.9024320 1.0866756
ever_married                 1.3416145 1.3104797 1.3734891
> |
```