Homework #1 - Joe Risi

Question 1

1. I filter out all missing observations. I go from 12144 records to 10200 records.
2. Recode family income into roughly equal sizes. The new categories are as follows:
   - 0 - 9,999
   - 10,000 - 19,999
   - 20,000 - 24,999
   - 25,000 - 34,999
   - 35,000 - 49,999
   - 50,000 - 74,999
   - 75,000 and above
3. I turn sex, byfaminc, and bys45 into dummy variables.
4. I drop the following categories which will serve as the reference categories for the regression:
   - female (sex)
   - 75,000 and above (byfaminc)
   - higher.sch.aftr.coll (bys45)

```
##
## Call:
## lm(formula = bygrads ~ ., data = dataWide)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.81319 -0.41509  0.05336  0.48159  2.00395
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            3.51841    0.02306 152.560  < 2e-16 ***
## won.t.finish.h.s      -1.17818    0.06214 -18.959  < 2e-16 ***
## will.finish.h.s       -0.92871    0.02587 -35.903  < 2e-16 ***
## voc.trd.bus.aftr.h.s -0.65136    0.02593 -25.115  < 2e-16 ***
## will.attend.college   -0.60532    0.02253 -26.870  < 2e-16 ***
## will.finish.college   -0.24041    0.01626 -14.788  < 2e-16 ***
## Less.than..10.000     -0.34418    0.02976 -11.567  < 2e-16 ***
## `10.000....19.999`    -0.22947    0.02747  -8.355  < 2e-16 ***
## `20.000..24.999`      -0.16325    0.02952  -5.530 3.28e-08 ***
## `25.000..34.999`      -0.15089    0.02622  -5.754 8.95e-09 ***
## `35.000..49.999`      -0.10332    0.02552  -4.049 5.19e-05 ***
## `50.000..74.999`      -0.06352    0.02715  -2.340   0.0193 *
## male                  -0.10190    0.01285  -7.932 2.38e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6448 on 10187 degrees of freedom
## Multiple R-squared:  0.2226, Adjusted R-squared:  0.2217
## F-statistic: 243.1 on 12 and 10187 DF,  p-value: < 2.2e-16
```

- All results are significant at the 0.05 level. All except one ($50,000 - %74,999) are significant at the

0.001 level.

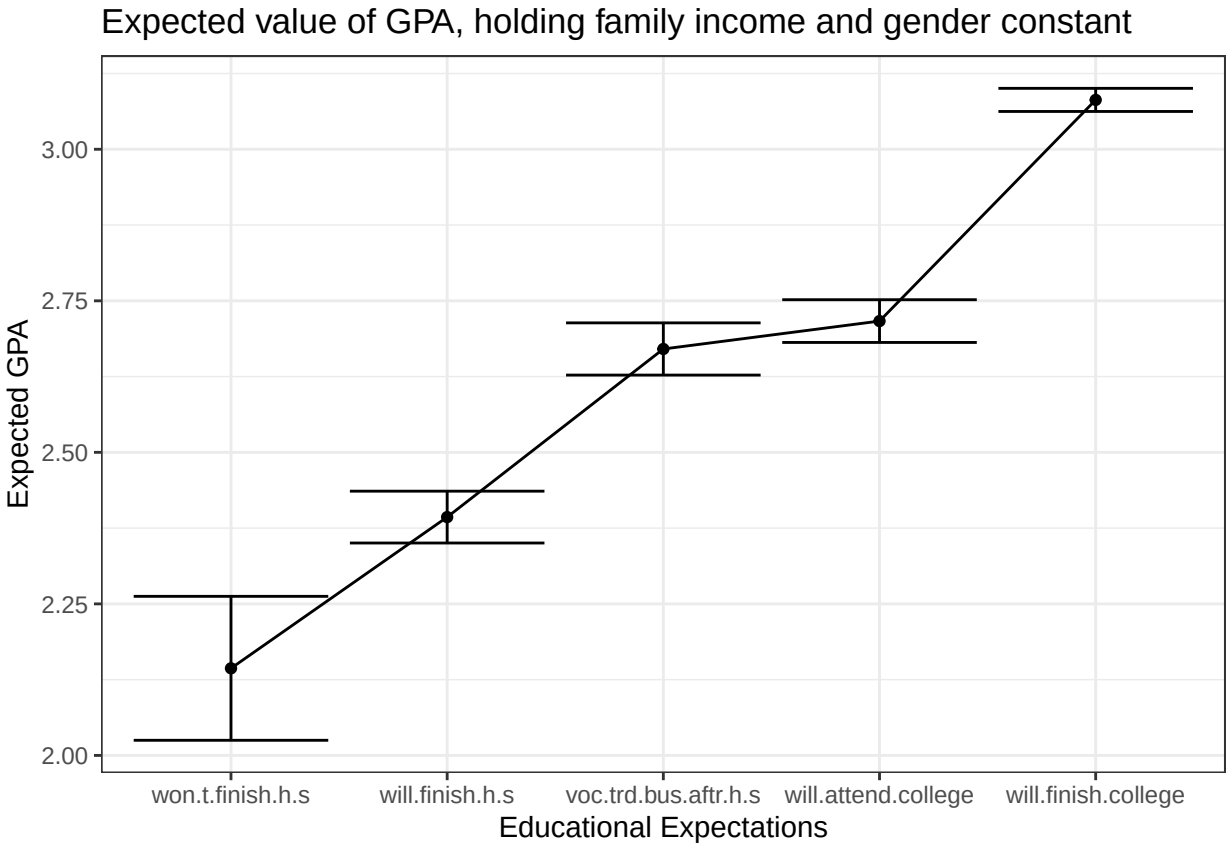
- Relative to females while holding all other variables in the model constant, being male decreases one's GPA by about 0.1 points, on average.
- Relative to those students who come from families making more $75,000 or more each year while holding all other variables in the model constant:
  - Coming from a family making less than $10,000 decreases your GPA by 0.34418 points on average.
  - Coming from a family making between $10,000 - $19,999 decreases your GPA by 0.22947 points on average.
  - Coming from a family making between $20,000 - $24,999 decreases your GPA by 0.16325 points on average.
  - Coming from a family making between $25,000 - $34,999 decreases your GPA by 0.15089 points on average.
  - Coming from a family making between $35,000 - $49,999 decreases your GPA by 0.10332 points on average.
  - Coming from a family making between $50,000 - $74,999 decreases your GPA by 0.06352 points on average.
- Relative to those students who have expectations of going beyond their college education while holding all other variables in the model constant:
  - Having expectations of not finishing high school decreases your GPA by 1.17818 points on average.
  - Having expectations of just finishing high school decreases your GPA by 0.92871 points on average.
  - Having expectations of going to vocational/trade school decreases your GPA by 0.65136 points on average.
  - Having expectations of attending college decreases your GPA by 0.60532 points on average.
  - Having expectations of finishing college decreases your GPA by 0.24041 points on average.

```
##
## Call:
## lm(formula = c("bygrads.z ~ won.t.finish.h.s + will.finish.h.s + voc.trd.bus.aftr.h.s + ",
## "    will.attend.college + will.finish.college + Less.than..10.000 + ",
## "    `10.000....19.999` + `20.000..24.999` + `25.000..34.999` + ",
## "    `35.000..49.999` + `50.000..74.999` + male"), data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8492 -0.5679  0.0730  0.6589  2.7419
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           0.72593    0.03156  23.005  < 2e-16 ***
## won.t.finish.h.s     -1.61205    0.08503 -18.959  < 2e-16 ***
## will.finish.h.s      -1.27070    0.03539 -35.903  < 2e-16 ***
## voc.trd.bus.aftr.h.s -0.89122    0.03549 -25.115  < 2e-16 ***
## will.attend.college  -0.82823    0.03082 -26.870  < 2e-16 ***
## will.finish.college  -0.32894    0.02224 -14.788  < 2e-16 ***
## Less.than..10.000    -0.47093    0.04071 -11.567  < 2e-16 ***
## `10.000....19.999`   -0.31397    0.03758  -8.355  < 2e-16 ***
## `20.000..24.999`     -0.22336    0.04039  -5.530 3.28e-08 ***
## `25.000..34.999`     -0.20646    0.03588  -5.754 8.95e-09 ***
## `35.000..49.999`     -0.14137    0.03492  -4.049 5.19e-05 ***
## `50.000..74.999`     -0.08691    0.03715  -2.340   0.0193 *
## male                 -0.13942    0.01758  -7.932 2.38e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.8822 on 10187 degrees of freedom
## Multiple R-squared:  0.2226, Adjusted R-squared:  0.2217
## F-statistic: 243.1 on 12 and 10187 DF,  p-value: < 2.2e-16
```

The above results represent **y-standardized** coefficients. It does not make sense to standardize my independent variables because they are all dummy variables. It's hard to interpret what a standard deviation in a dummy variable would mean.

- All results are significant at the 0.05 level. All except one ($50,000 - %74,999) are significant at the 0.001 level.
- Relative to females while holding all other variables in the model constant, being male decreases one's GPA by about 0.13942 standard deviations on average.
- Relative to those students who come from families making more $75,000 or more each year while holding all other variables in the model constant:
  - Coming from a family making less than $10,000 decreases your GPA by 0.47093 standard deviations.
  - Coming from a family making between $10,000 - $19,999 decreases your GPA by 0.31397 standard deviations on average.
  - Coming from a family making between $20,000 - $24,999 decreases your GPA by 0.22336 standard deviations on average.
  - Coming from a family making between $25,000 - $34,999 decreases your GPA by 0.20646 standard deviations on average.
  - Coming from a family making between $35,000 - $49,999 decreases your GPA by 0.14137 standard deviations on average.
  - Coming from a family making between $50,000 - $74,999 decreases your GPA by 0.08691 standard deviations on average.
- Relative to those students who have expectations of going beyond their college education while holding all other variables in the model constant:
  - Having expectations of not finishing high school decreases your GPA by 1.61205 standard deviations on average.
  - Having expectations of just finishing high school decreases your GPA by 1.27070 standard deviations on average.
  - Having expectations of going to vocational/trade school decreases your GPA by 0.89122 standard deviations on average.
  - Having expectations of attending college decreases your GPA by 0.82823 standard deviations on average.
  - Having expectations of finishing college decreases your GPA by 0.32894 standard deviations on average.

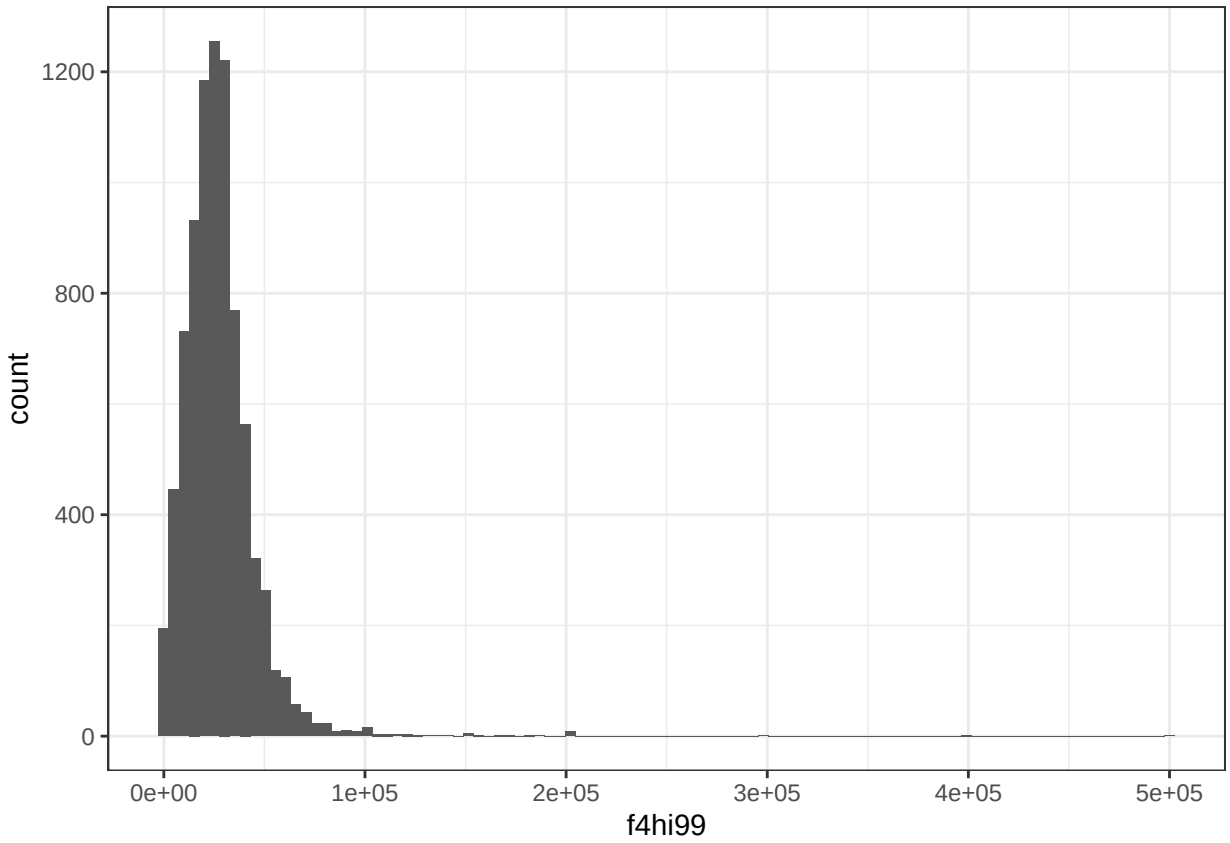


By holding family income and sex constant, I held each categorical variable at its mean value.

Question 2

1. I filter out all missing observations and all observations where the individual made $0. I go from 12144 records to 8341 records.
    - Another possible data cleaning decision could have been to drop all observations where f4aempl == 0 (the individual was not employed for pay). It seemed important to me to capture the fact that some of these individuals still reported income (there are some individual who reportedly didn't work but still reported income), and I do not drop them.
    - It is sometimes the case that incomes 5 times above the median or greater are dropped. I keep them and see if any of them show up as outliers later.
    - It can be also be common to practice to log transform the income variable. This is done to account for the right-hand skew of the data. However the data looks pretty normal (very few outlier values, see plot below) so I did not log transform the income variable.
2. I turn partnership status (f4gmrs) into a series of dummy variables, and I drop the single, never married category so it can serve as the reference.

```
## Don't know how to automatically pick scale for object of type haven_labelled/vctrs_vctr/double. Defau
```



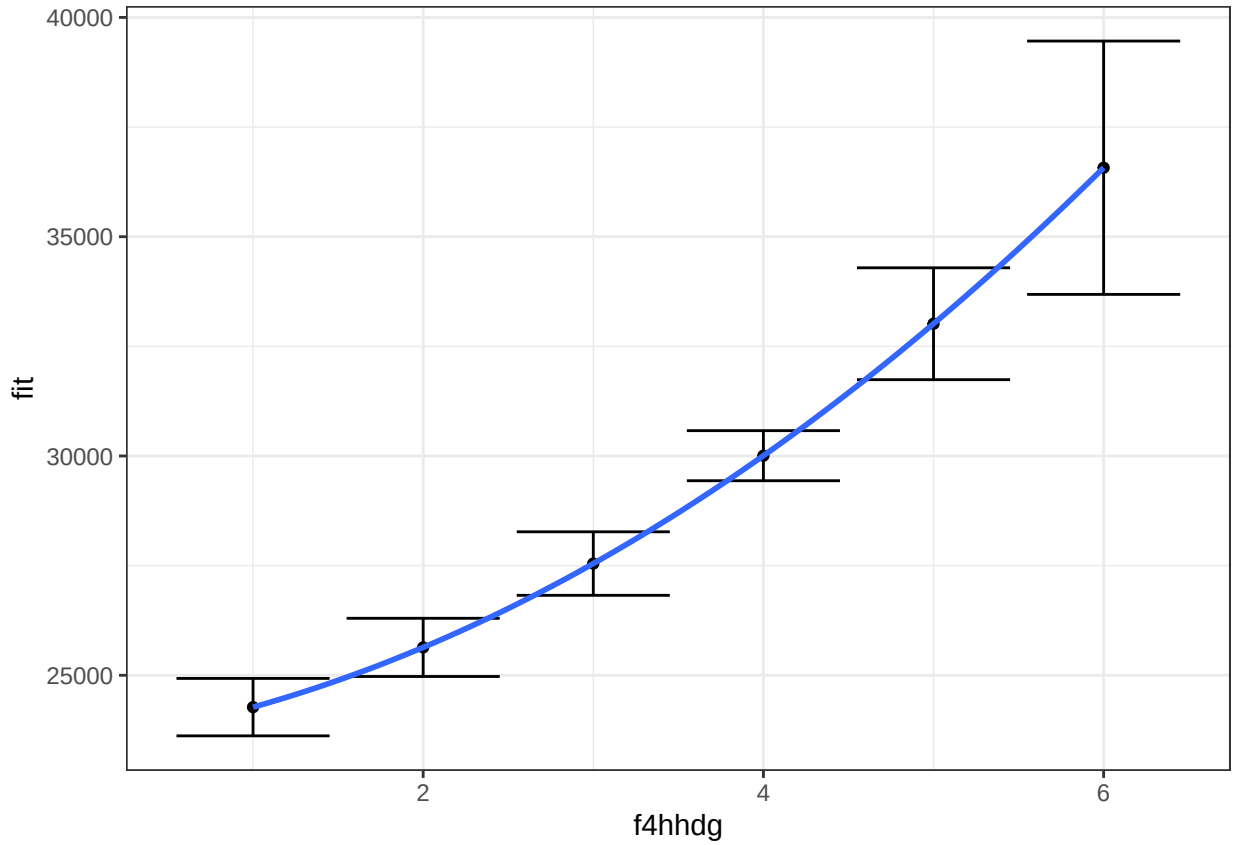

```
##
## Call:
## lm(formula = f4hi99 ~ f4hhdg, data = dataClean2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -32094 -10090  -2083   5917 469910
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22081.0      436.7   50.56   <2e-16 ***
## f4hhdg        2002.2      145.1   13.80   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19120 on 8339 degrees of freedom
## Multiple R-squared:  0.02232,    Adjusted R-squared:  0.0222
## F-statistic: 190.4 on 1 and 8339 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = f4hi99 ~ f4hhdg + f4hhdg_squared, data = dataClean2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -34573 -10047  -2007   6364 469993
##
```

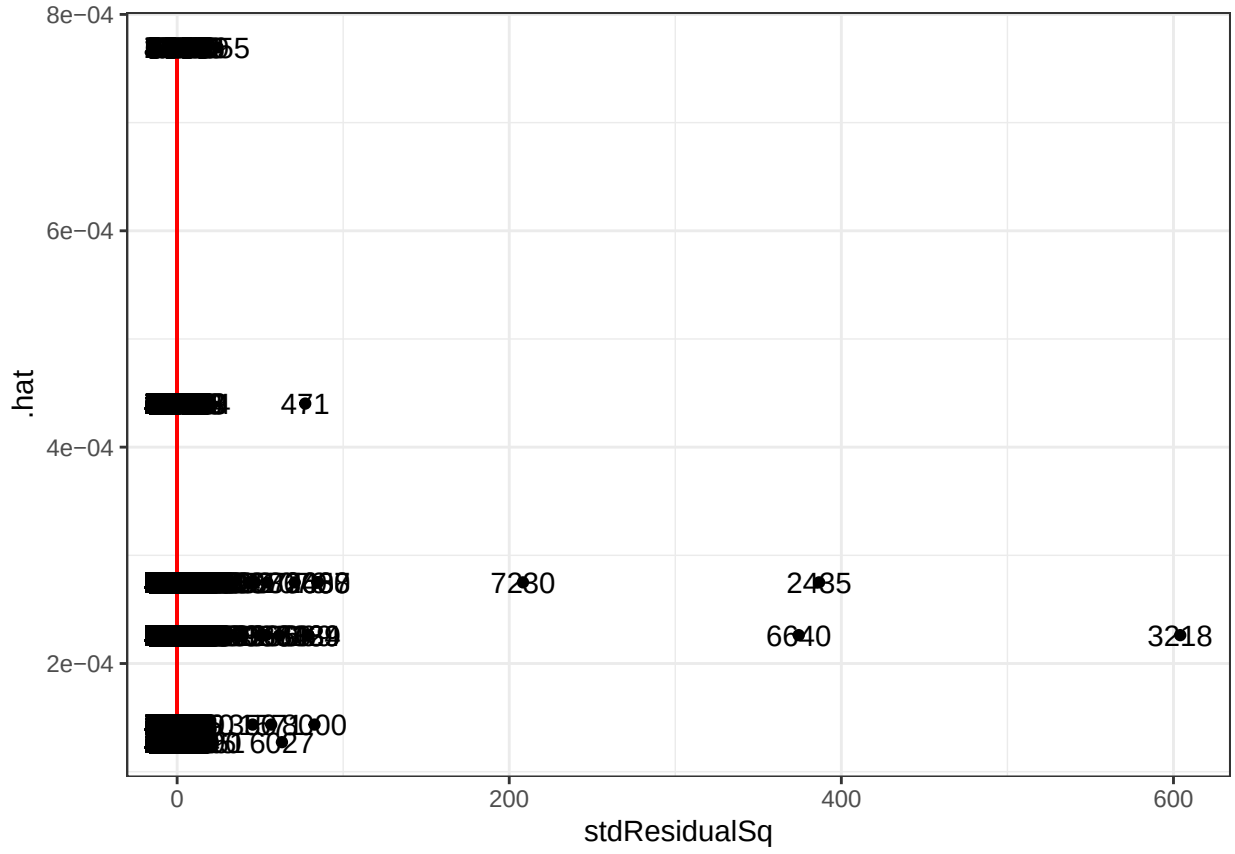
```
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     23459.7      880.5  26.645   <2e-16 ***
## f4hhdg            539.2      824.2   0.654   0.5130
## f4hhdg_squared    274.4      152.2   1.803   0.0714 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19120 on 8338 degrees of freedom
## Multiple R-squared:  0.0227, Adjusted R-squared:  0.02247
## F-statistic: 96.85 on 2 and 8338 DF,  p-value: < 2.2e-16
```

```
## Analysis of Variance Table
##
## Model 1: f4hi99 ~ f4hhdg
## Model 2: f4hi99 ~ f4hhdg + f4hhdg_squared
##   Res.Df        RSS Df  Sum of Sq      F  Pr(>F)
## 1   8339 3.0490e+12
## 2   8338 3.0478e+12  1 1188618865 3.2518 0.07138 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is marginal evidence to suggest the effect of educational attainment on yearly income may be non-linear. A plot may prove useful.



The graphical evidence suggests there may be a slight curve in the data with increasing returns to education (each extra unit of education produces a bigger bump in income than the previous unit increase). The evidence though is suggestive and not definitive.




```
##
## Call:
## lm(formula = f4hi99 ~ f4hhdg, data = dataClean2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -32094 -10090  -2083   5917 469910
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22081.0      436.7   50.56   <2e-16 ***
## f4hhdg        2002.2      145.1   13.80   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19120 on 8339 degrees of freedom
## Multiple R-squared:  0.02232,    Adjusted R-squared:  0.0222
## F-statistic: 190.4 on 1 and 8339 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = f4hi99 ~ f4hhdg, data = noOutliers)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -31804  -9904  -1854   6171 176072
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  21953.1      391.3   56.10   <2e-16 ***
## f4hhdg        1975.2      130.0   15.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17130 on 8334 degrees of freedom
## Multiple R-squared:  0.02694,    Adjusted R-squared:  0.02682
## F-statistic: 230.7 on 1 and 8334 DF,  p-value: < 2.2e-16
```

There are some potential outliers but removing them from the linear regression doesn't change the model very much so the outliers aren't concerning.

```
##
## Call:
## lm(formula = f4hi99 ~ ., data = dataClean2a)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -34313 -10213  -1949   6316 469323
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     23170.4      903.5  25.646   <2e-16 ***
## f4hhdg                            504.4      824.1   0.612   0.5405
## f4hhdg_squared                    281.0      152.3   1.846   0.0650 .
## divorced                        -1285.6     1124.9  -1.143   0.2531
## in.marriage.like.relationship    4097.0     2255.2   1.817   0.0693 .
## married                           993.1      443.8   2.238   0.0253 *
## separated                       -3453.0     2227.0  -1.551   0.1211
## widowed                        -21303.2    19112.1  -1.115   0.2650
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19110 on 8333 degrees of freedom
## Multiple R-squared:  0.02439,    Adjusted R-squared:  0.02357
## F-statistic: 29.77 on 7 and 8333 DF,  p-value: < 2.2e-16
```

```
## Analysis of Variance Table
##
## Model 1: f4hi99 ~ f4hhdg + f4hhdg_squared
## Model 2: f4hi99 ~ f4hhdg + f4hhdg_squared + divorced + in.marriage.like.relationship +
##     married + separated + widowed
##   Res.Df        RSS Df  Sum of Sq     F  Pr(>F)
## 1   8338 3.0478e+12
## 2   8333 3.0425e+12  5 5274117141 2.889 0.01307 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA test indicates a statistically significant amount of more variation is explained when including the dummy variables for partnership status. However only a few of the partnership dummy variables are statistically significant, and the differences in the R^2 values between the two models is very small. Substantively these variables don't add much to the model.

- The coefficient for f4hhdg is a bit contextual as we're now stating the effect of education on income is

dependent upon what specific level of education one has. It shall be interpreted below.

- The coefficient for the f4hhdg_squared term is positive suggesting the curve is convex meaning the more education one gets, the higher the boost in earnings one gets for each extra level of education obtained (statistically significant at the 0.1 level).
- If we wanted to get a sense for how much getting an Associate's Degree adds to your income, we can take the derivative:
  - Constant Effect of Education (fhddg) ~ 504.4
  - Changing Effect of Education (fhddg_squared) ~ 281.0 (5 for Associate's Degree, 2 for derivative term)
  - Boost in Income ~ 504.4 + 5 * 2 * 281 = 3314.4
  - Compare to the boost in income you get from having a Bachelor's Degree: 504.4 + 6 * 2 * 281 ~ 3876.4. Notice how the boost went up, consistent with our observation of the increasing returns of education.
- Relative to those who were single and never married as of 2000 while holding all other variables in the model constant:
  - Being divorced decreases your income by $1285.6 on average.
  - Being in a marriage-like relationship increases your income by $4097.0 on average (statistically significant at the 0.1 level).
  - Being in a marriage increases your income by $993.1 on average (statistically significant at the 0.05 level).
  - Being separated decreases your income by $3453.0 on average.
  - Being widowed decreases your income by $21303.2 on average (only 1 person in the sample was widowed).
- What is the expected yearly income of a respondent who is married with an Associate's Degree?
  - Intercept ~ 23170.40
  - Married ~ 993.10
  - Constant Effect of Education (fhddg) ~ 504.4
  - Changing Effect of Education (fhddg_squared) ~ 281.0
  - 23170.40 + 993.10 + 5 * 504.4 + 281 * 25 ~ 33710.46

Question 3

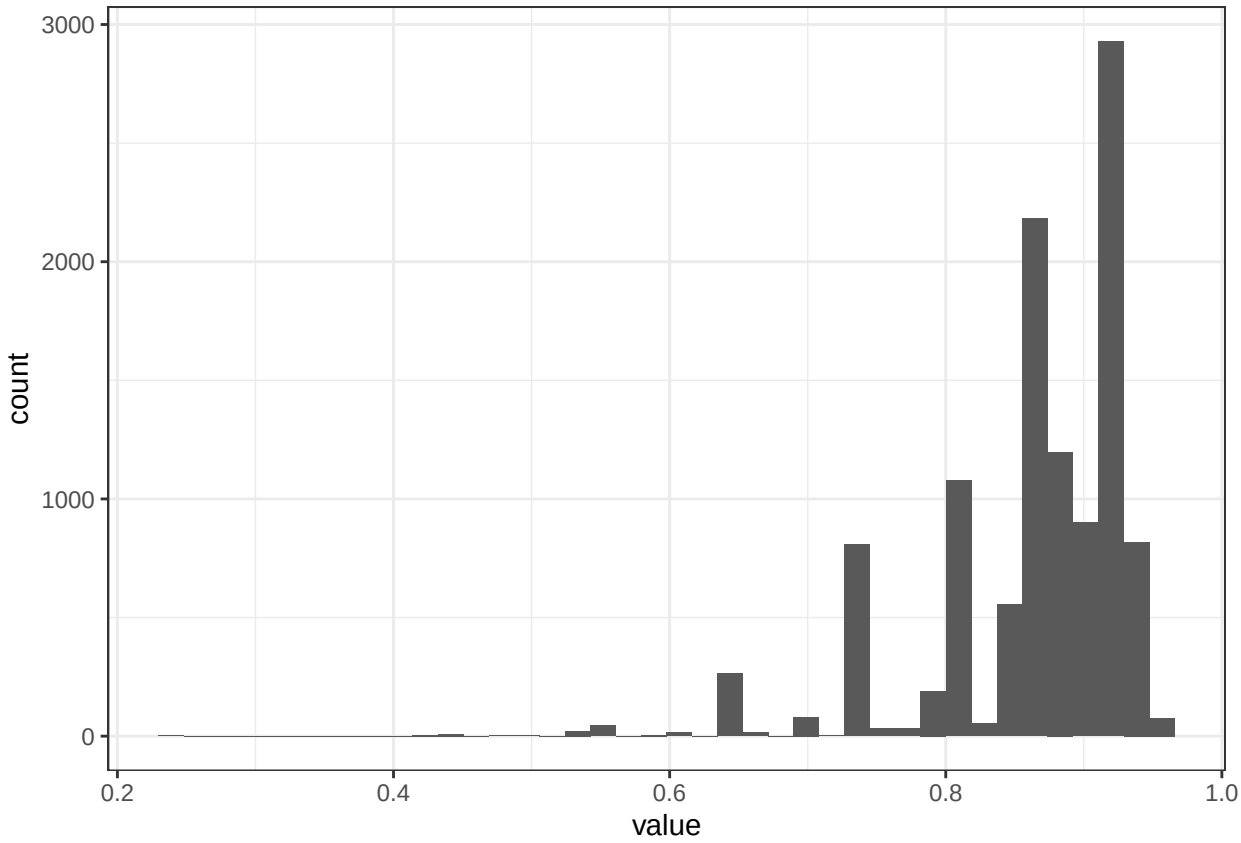
1. I filter out all missing observations. I go from 12144 records to 11331 records.
2. I turn partnership status (f4gmrs) and sex (sex) into dummy variables. I drop males and those who were single, never married to serve as the reference groups.

```
##
## Call:
## glm(formula = f4aempl ~ ., family = "binomial", data = dataWide3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4288   0.3840   0.4739   0.5344   1.5987
##
## Coefficients:
##                                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)                          2.57022    0.05612  45.796   <2e-16 ***
## f4gnch                              -0.44019    0.02720 -16.182   <2e-16 ***
## f4gmrs.divorced                      0.32549    0.13796   2.359   0.0183 *
## f4gmrs.in.marriage.like.relationship -0.16750    0.27123  -0.618   0.5369
## f4gmrs.married                       0.06430    0.06289   1.022   0.3066
## f4gmrs.separated                     0.13084    0.22014   0.594   0.5523
## f4gmrs.widowed                       0.33695    1.16215   0.290   0.7719
## sex.female                          -0.69595    0.06032 -11.537   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8988.9  on 11330  degrees of freedom
## Residual deviance: 8510.8  on 11323  degrees of freedom
## AIC: 8526.8
##
## Number of Fisher Scoring iterations: 5
```

```
##                                f4gnch                       f4gmrs.divorced
##                           -35.608812                             38.470745
## f4gmrs.in.marriage.like.relationship                        f4gmrs.married
##                           -15.421981                             6.641545
##                      f4gmrs.separated                        f4gmrs.widowed
##                             13.978478                             40.066759
##                           sex.female
##                           -50.140095
```

- The number of children is highly statistically significant as is being female. The relationship variables aren't really statistically significant except for being divorced (significant at the 0.05 level).
- Holding all other variables in the model constant, each additional child decreases the odds of being employed by 35.6% on average.
- Holding all other variables in the model constant, being female decreases the odds of being employed by 50.1% on average relative to males.
- Relative to people who are single and have never been married while holding all other variables in the model constant:
  - Being divorced increases the odds of being employed by 38.5% on average.
  - Being in a marriage-like relationship decreases the odds of being employed by 15.4% on average.
  - Being married increases the odds of being employed by 6.6% on average.
  - Being separated increases the odds of being employed by 14.0% on average.
  - Being widowed increases the odds of being employed by 40.1% on average.



Above are the predicted probabilities. As you can see, most people are predicted as being employed.

- The predicted probability of being employed for the average respondent is ~0.8778169.
- The predicted probability of being employed for a single mother with exactly 3 children is ~0.6349911.
- The predicted probability of being employed as a married father is ~ 0.9140363.