Visualizing Generalized Linear Models and Generalized Additive Models

Pauline I. Alvarado

Spring 2018

This coding exercise was from the Program Evaluation and Data Analysis course taught by Dr. Nelson Lim at the University of Pennsylvania. Data was provided by the instructor.

Data Management

Load data and packages

```
library(tidyverse)
library(ggplot2)
library(readstata13)
library(modelr)

acs <- read.dta13("acsphillylaborforce.dta")
```

Create duplicate viarables for better graphic & summary tables

```
acs <- acs %>% rename(Race = raceth,
                      Sex  = sex,
                      Education = educ_year,
                      Degree = education,
                      Occupation = gen_occ,
                      Industry = ind_cat5,
                      Income = incwage,
                      Managers = leader_cat,
                      Age = age,
                      College_major = major1,
                      Marital_status = marst) %>%
               mutate(Age = as.numeric(Age),
                      Age_sq = Age*Age,
                      Education11 = case_when((Education < 11) ~ 0,
                                              (Education >= 11) ~ Education - 11),
                      Education16 = case_when((Education < 16) ~ 0,
                                              (Education >= 16) ~ Education - 16))
```

Generalized Linear Models

Create linear models

```
# NOTE: family = gaussian (normal distribution)

fit1 <- glm(Income ~ Sex, family = gaussian(link = identity), data=acs)


fit2 <- glm(Income ~ Sex + Race, family = gaussian(link = identity), data = acs)

fit3 <- glm(Income ~ Sex + Race + Education + Education11 + Education16,
            family = gaussian(link = identity), data = acs)

fit4 <- glm(Income ~ Sex + Race + Education + Education11 + Education16 +
            Age + Age_sq, family = gaussian(link = identity), data = acs)

fit5 <- glm(Income ~ Sex + Race + Education + Education11 + Education16 +
            Age + Age_sq + Managers, family = gaussian(link = identity), data = acs)
```

Compare models

```
anova(fit1, fit5, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Income ~ Sex
## Model 2: Income ~ Sex + Race + Education + Education11 + Education16 +
##     Age + Age_sq + Managers
##   Resid. Df Resid. Dev Df   Deviance  Pr(>Chi)
## 1      3596 8.7289e+12
## 2      3583 6.6435e+12 13 2.0854e+12 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Visualize models in tables

```
library("texreg")
screenreg(list(fit1,fit2,fit3))
```

```
##
## ================================================================================
##                 Model 1               Model 2               Model 3
## --------------------------------------------------------------------------------
## (Intercept)             50405.72 ***          57574.61 ***          43999.72 ***
##                         (1215.12)             (1412.72)             (6957.28)
## SexFemale              -10165.54 ***          -8884.53 ***         -11650.77 ***
```

```
##                          (1648.88)             (1637.01)            (1532.44)
## RaceB-NH                                     -15856.86 ***         -5854.58 **
##                                               (1858.77)            (1797.20)
## RaceHispanic                                 -18957.18 ***         -6808.46 *
##                                               (3036.06)            (2902.68)
## RaceA-NH                                     -14176.56 ***        -14501.05 ***
##                                               (3157.36)            (2990.86)
## RaceAI-NH                                     22906.15             24601.12
##                                              (19892.93)           (18657.79)
## RaceOther                                    -12693.87 *           -9773.58 *
##                                               (5197.11)            (4848.99)
## Education                                                           -807.43
##                                                                     (640.54)
## Education11                                                         6200.66 ***
##                                                                     (883.80)
## Education16                                                         5820.21 ***
##                                                                    (1138.82)
## ---------------------------------------------------------------------------------
## AIC                      87967.77              87878.56             87366.53
## BIC                      87986.33              87928.06             87434.60
## Log Likelihood          -43980.88             -43931.28            -43672.26
## Deviance         8728917213593.88      8491523001555.09     7352866554995.64
## Num. obs.                   3598                  3598                 3598
## =================================================================================
## *** p < 0.001, ** p < 0.01, * p < 0.05
```

Use Packages to Automate Work: Spline Functions

Breaks down independent variables into a small number of segments connected by knots.

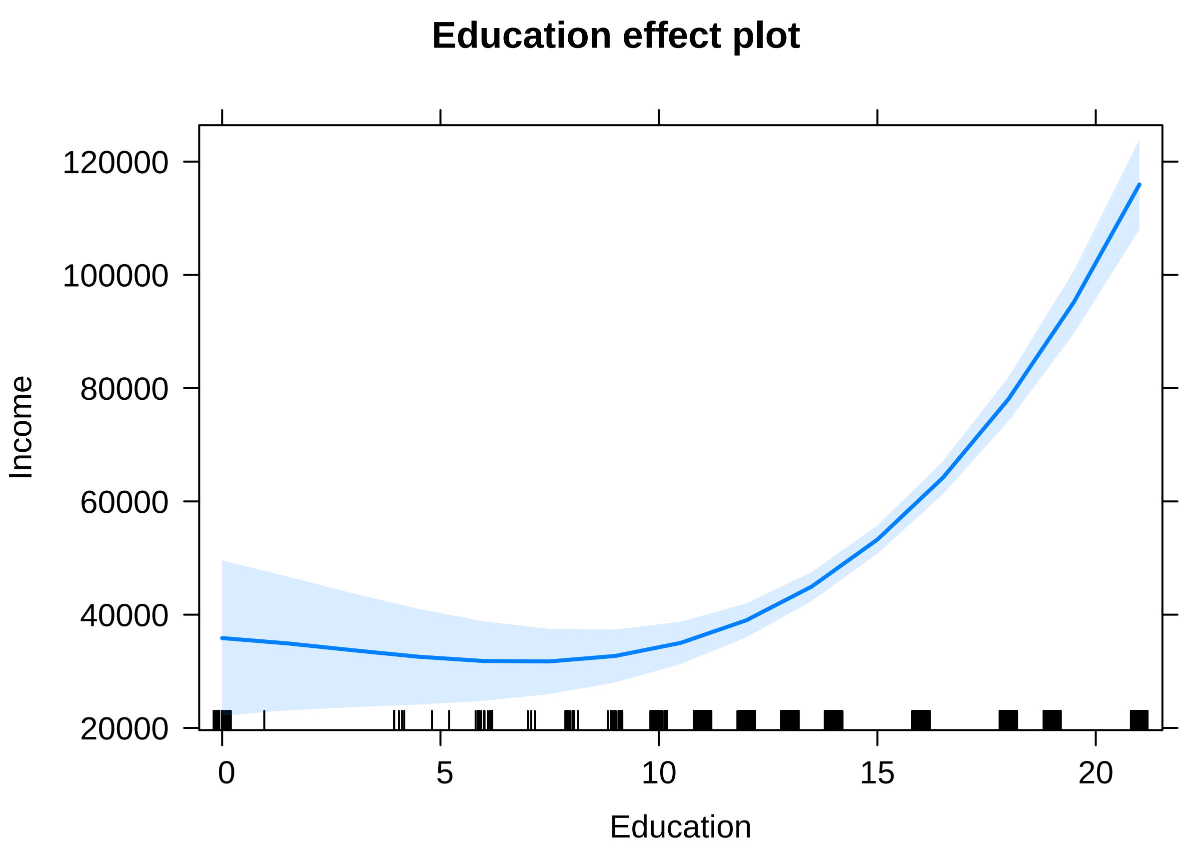
```r
library(splines)

fit6 <- glm(Income ~ Sex + Race + bs(Education,3) + poly(Age,2) + Managers,
            family = gaussian(link = identity), data = acs)
summary(fit6)
```

```
##
## Call:
## glm(formula = Income ~ Sex + Race + bs(Education, 3) + poly(Age,
##     2) + Managers, family = gaussian(link = identity), data = acs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -111736   -19082    -5131    11119   395761
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)            85378      15344   5.564 2.82e-08 ***
## SexFemale             -10271       1462  -7.024 2.56e-12 ***
## RaceB-NH               -7742       1716  -4.511 6.66e-06 ***
## RaceHispanic           -5750       2784  -2.065 0.038959 *
## RaceA-NH              -11576       2847  -4.066 4.89e-05 ***
## RaceAI-NH              16792      17779   0.944 0.344989
```

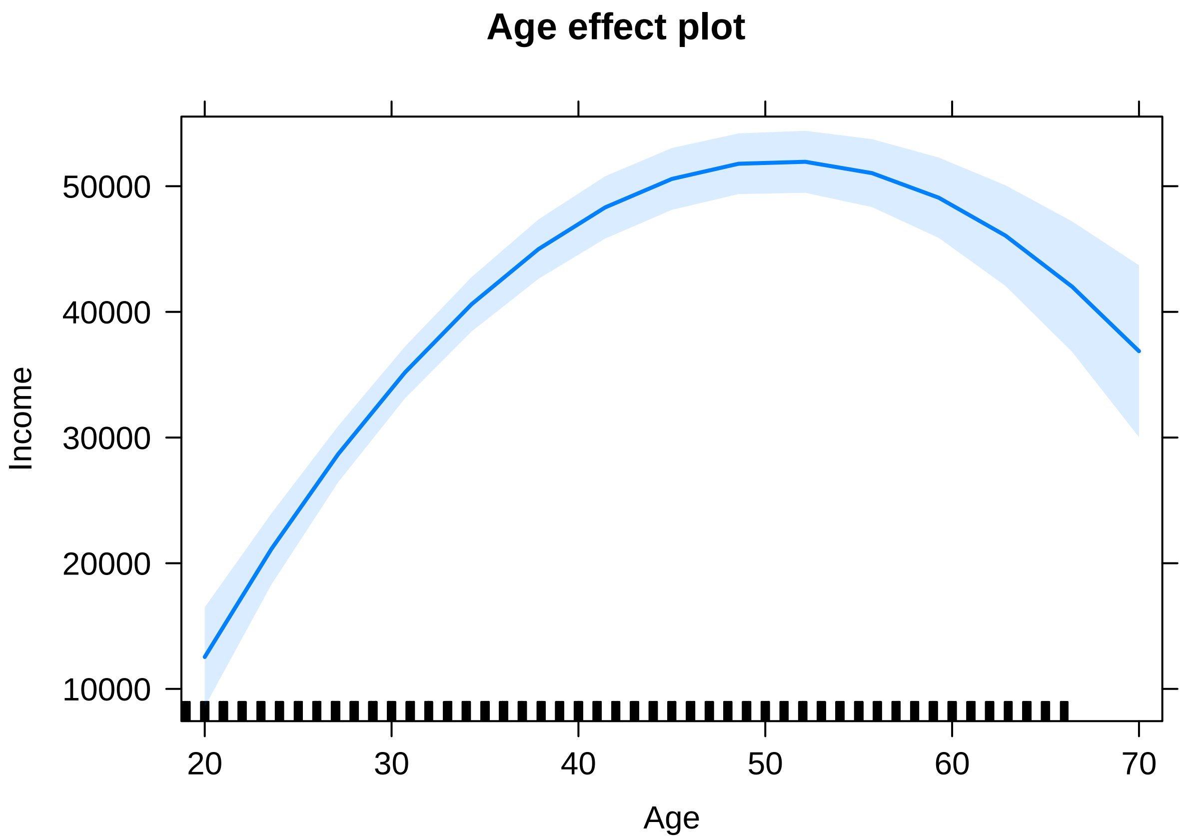
```
## RaceOther               -5678       4616  -1.230 0.218771
## bs(Education, 3)1       -3298      14768  -0.223 0.823292
## bs(Education, 3)2      -25652      10252  -2.502 0.012392 *
## bs(Education, 3)3       80095       8254   9.703  < 2e-16 ***
## poly(Age, 2)1          581973      44091  13.199  < 2e-16 ***
## poly(Age, 2)2         -377024      43840  -8.600  < 2e-16 ***
## ManagersManagers       -26009      13875  -1.874 0.060945 .
## ManagersSupervisors    -33677      14113  -2.386 0.017077 *
## ManagersOther          -49980      13682  -3.653 0.000263 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 1853472516)
##
##     Null deviance: 8.8212e+12  on 3597  degrees of freedom
## Residual deviance: 6.6410e+12  on 3583  degrees of freedom
## AIC: 87010
##
## Number of Fisher Scoring iterations: 2
```

Create plots

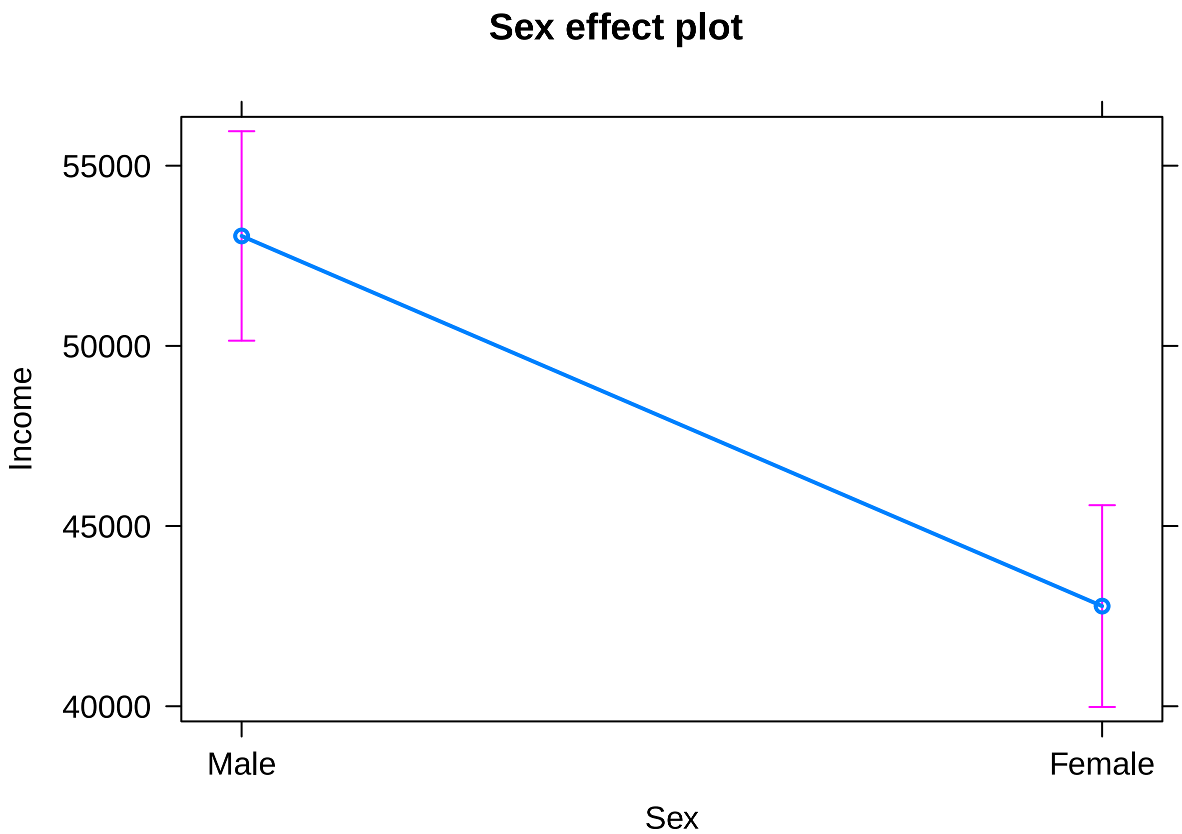
```r
library(effects)

plot(effect("bs(Education,3)", fit6))
```



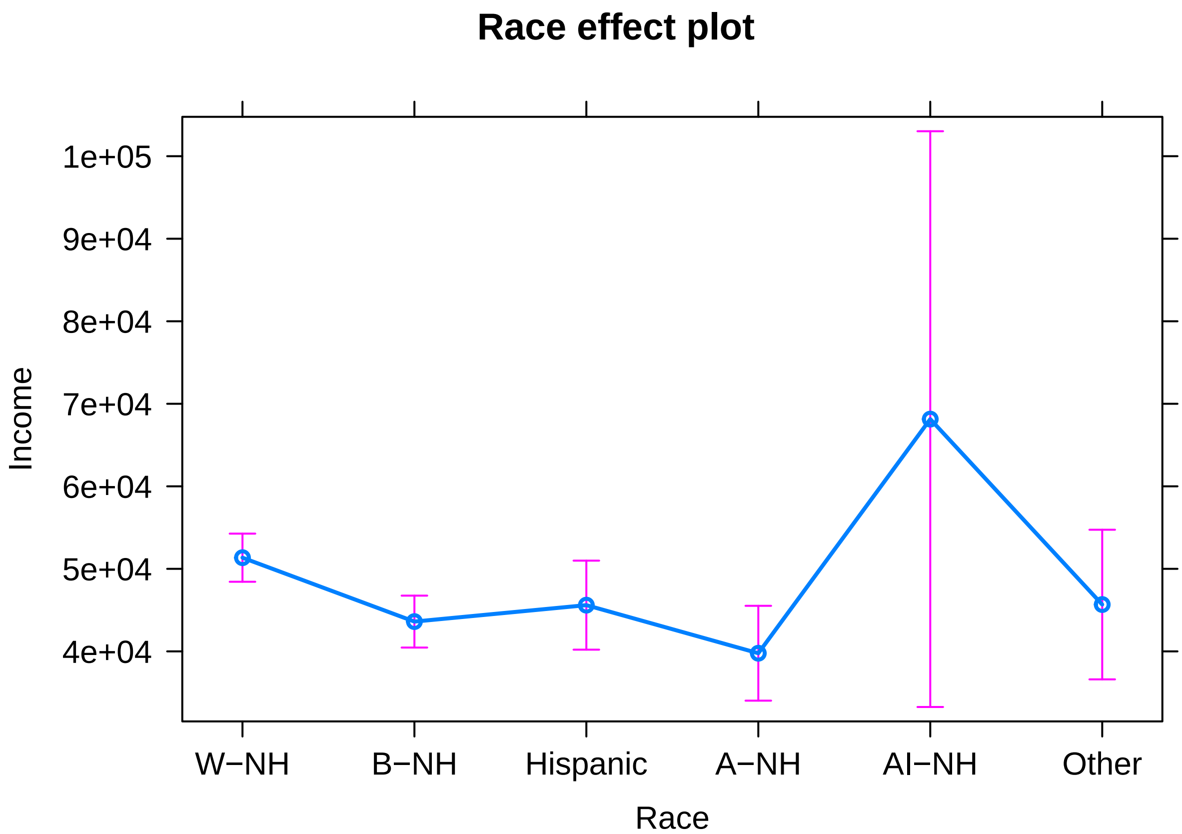
```r
plot(effect("poly(Age,2)", fit6))
```



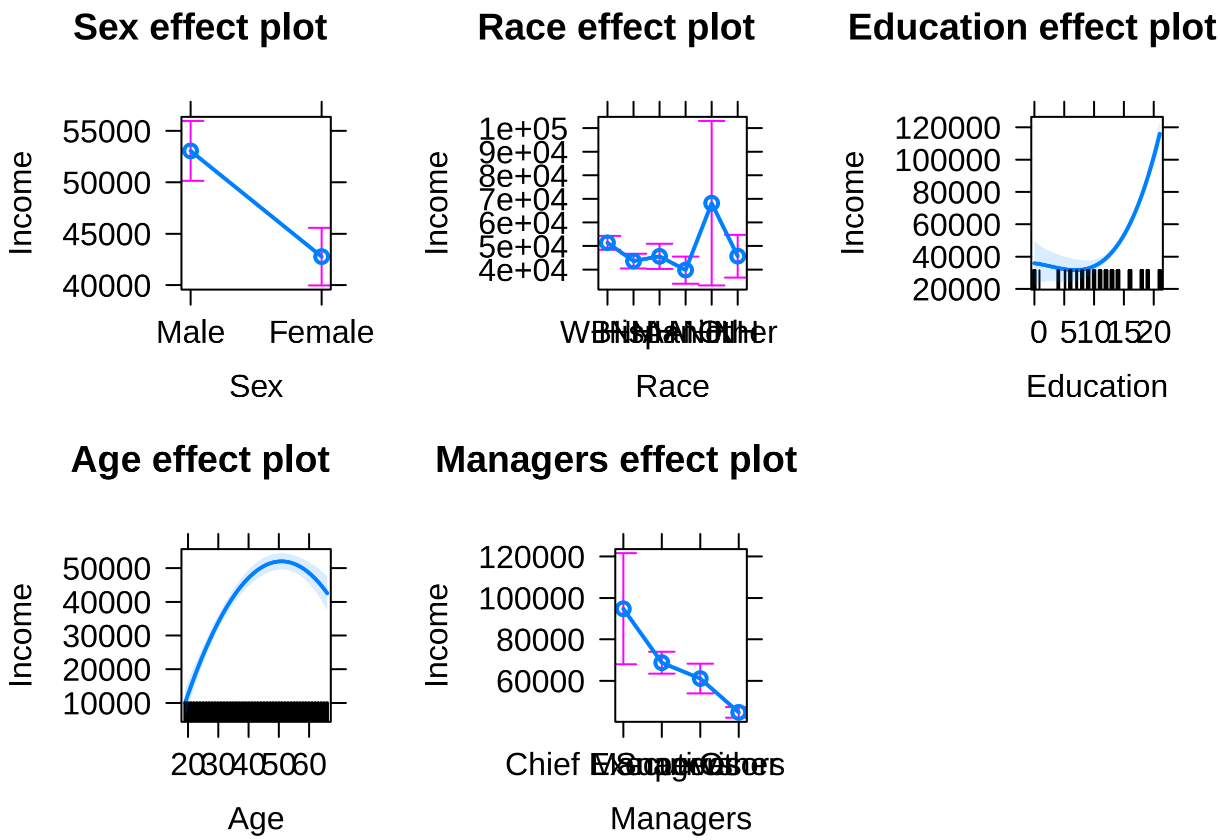

```r
plot(effect("Sex", fit6))
```




```
plot(effect("Race", fit6))
```




```
plot(allEffects(fit6, xlevels = 50))
```



Generalized Additive Model

Only assume variables are additive and not linear. More flexible regression models. ## Create models and view summaries

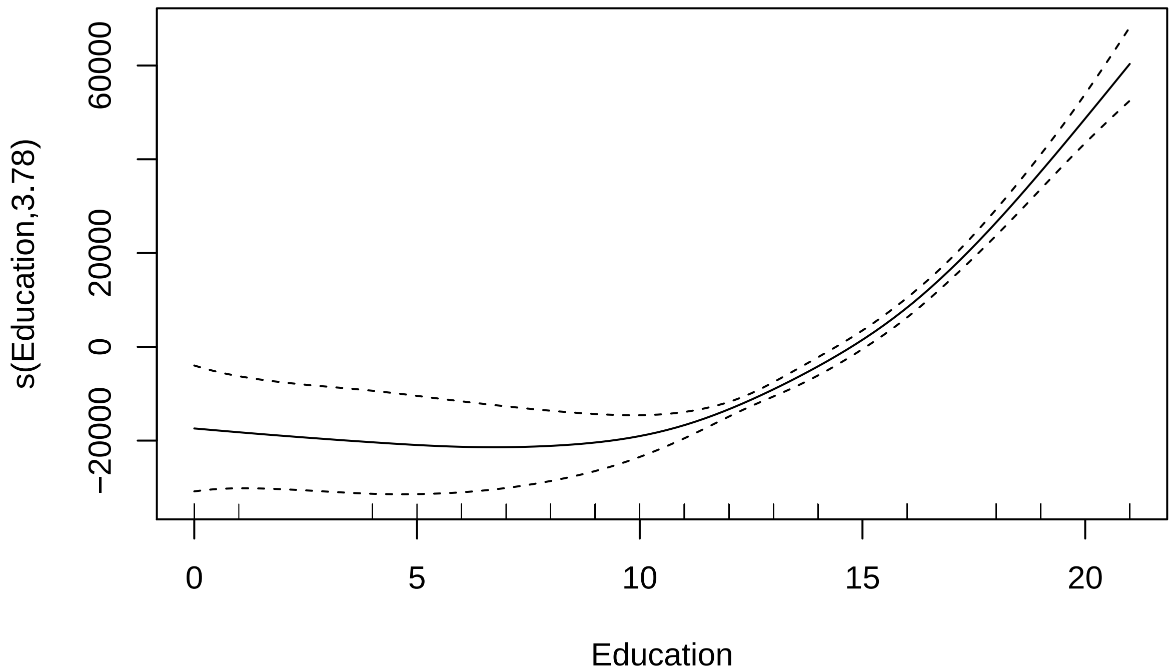
```
library(mgcv)
    fit_gam1 <- gam(Income ~ Sex + Race + Education + Age + Managers, data = acs)
    summary(fit_gam1)

    fit_gam2 <- gam(Income ~ Sex + Race + s(Education) + s(Age) + Managers, data = acs)
    summary(fit_gam2)
```

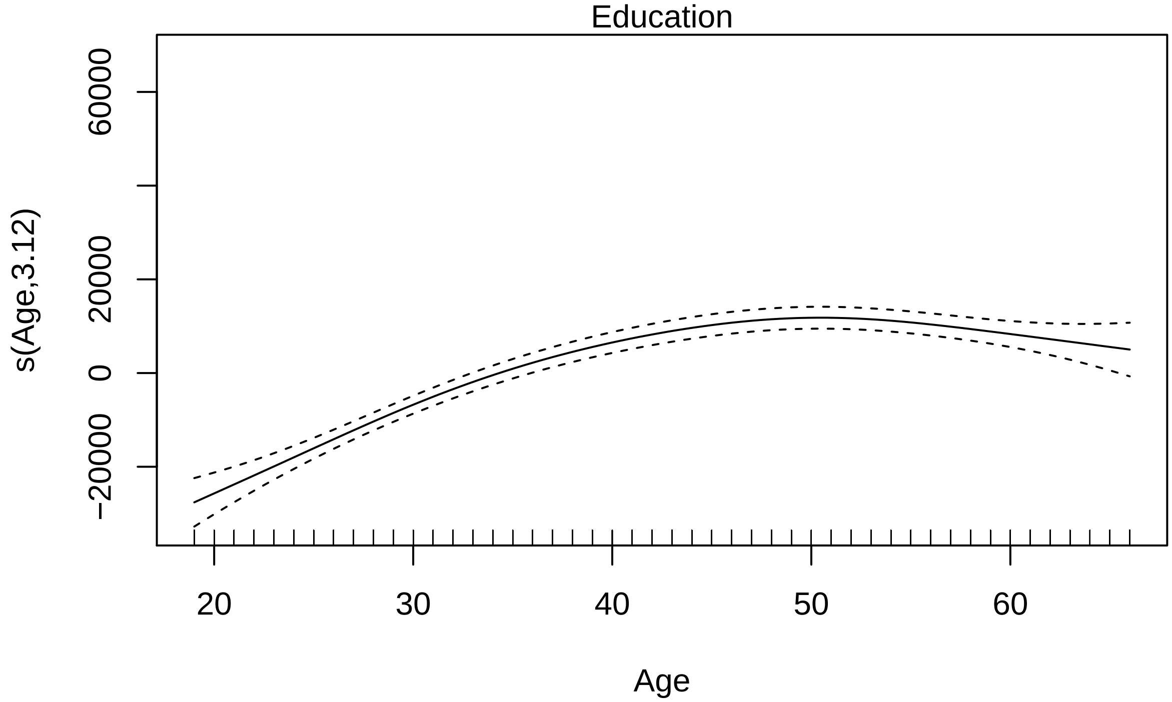
```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Income ~ Sex + Race + Education + Age + Managers
##
## Parametric coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -2951.60   14897.40  -0.198 0.842956
## SexFemale           -10681.34    1499.71  -7.122 1.28e-12 ***
## RaceB-NH             -8420.16    1748.89  -4.815 1.54e-06 ***
## RaceHispanic         -3498.02    2841.43  -1.231 0.218374
```

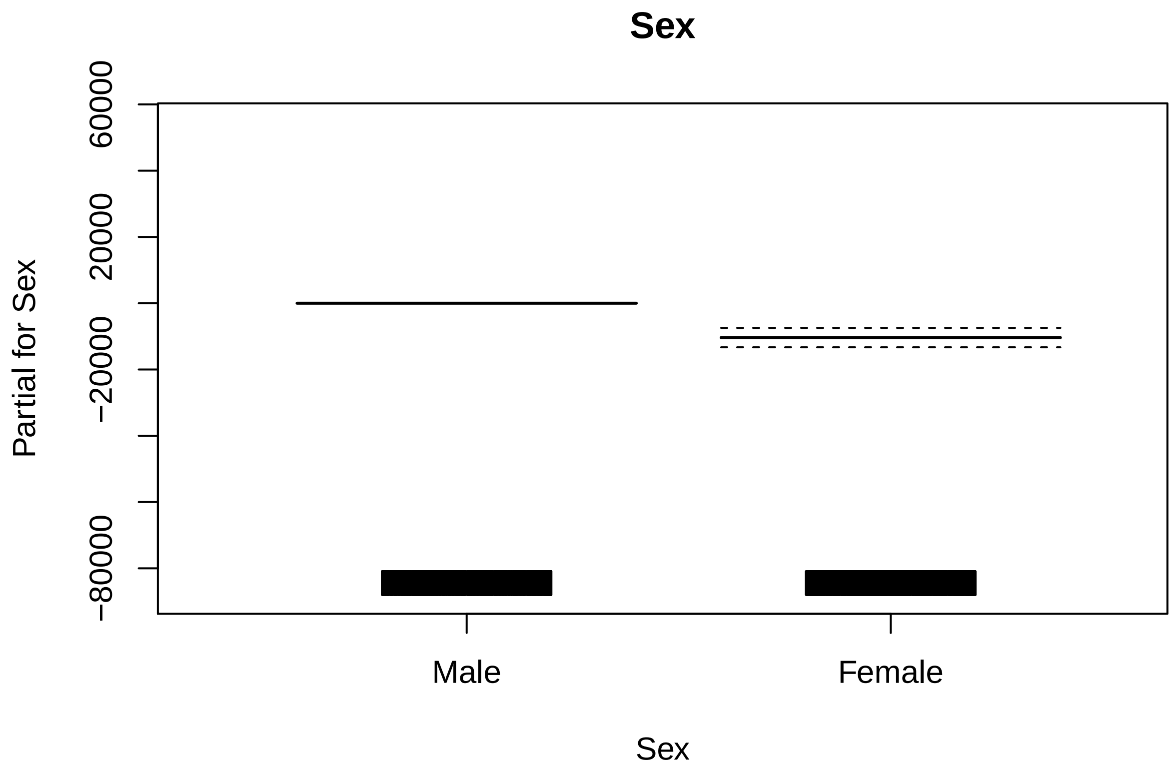
```
## RaceA-NH             -6382.84    2894.04  -2.206 0.027482 *
## RaceAI-NH            37276.51   18186.97   2.050 0.040473 *
## RaceOther            -3607.07    4748.53  -0.760 0.447532
## Education             5372.50     257.29  20.881  < 2e-16 ***
## Age                    794.80      56.77  14.000  < 2e-16 ***
## ManagersManagers    -28596.67   14281.05  -2.002 0.045315 *
## ManagersSupervisors -36492.69   14522.38  -2.513 0.012019 *
## ManagersOther       -53037.32   14074.40  -3.768 0.000167 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## R-sq.(adj) =  0.199   Deviance explained = 20.2%
## GCV = 1.9707e+09  Scale est. = 1.9641e+09  n = 3598
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Income ~ Sex + Race + s(Education) + s(Age) + Managers
##
## Parametric coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           101615      13648   7.446 1.20e-13 ***
## SexFemale             -10378       1462  -7.096 1.54e-12 ***
## RaceB-NH               -7695       1722  -4.468 8.13e-06 ***
## RaceHispanic           -5546       2784  -1.992 0.046430 *
## RaceA-NH              -11438       2848  -4.016 6.03e-05 ***
## RaceAI-NH              17275      17778   0.972 0.331279
## RaceOther              -5644       4620  -1.222 0.221857
## ManagersManagers      -26067      13876  -1.879 0.060384 .
## ManagersSupervisors   -33605      14115  -2.381 0.017326 *
## ManagersOther         -49956      13682  -3.651 0.000265 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                edf Ref.df     F p-value
## s(Education) 3.779   4.63 111.8  <2e-16 ***
## s(Age)       3.121   3.90  62.8  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.244   Deviance explained = 24.7%
## GCV = 1.8624e+09  Scale est. = 1.8536e+09  n = 3598
```

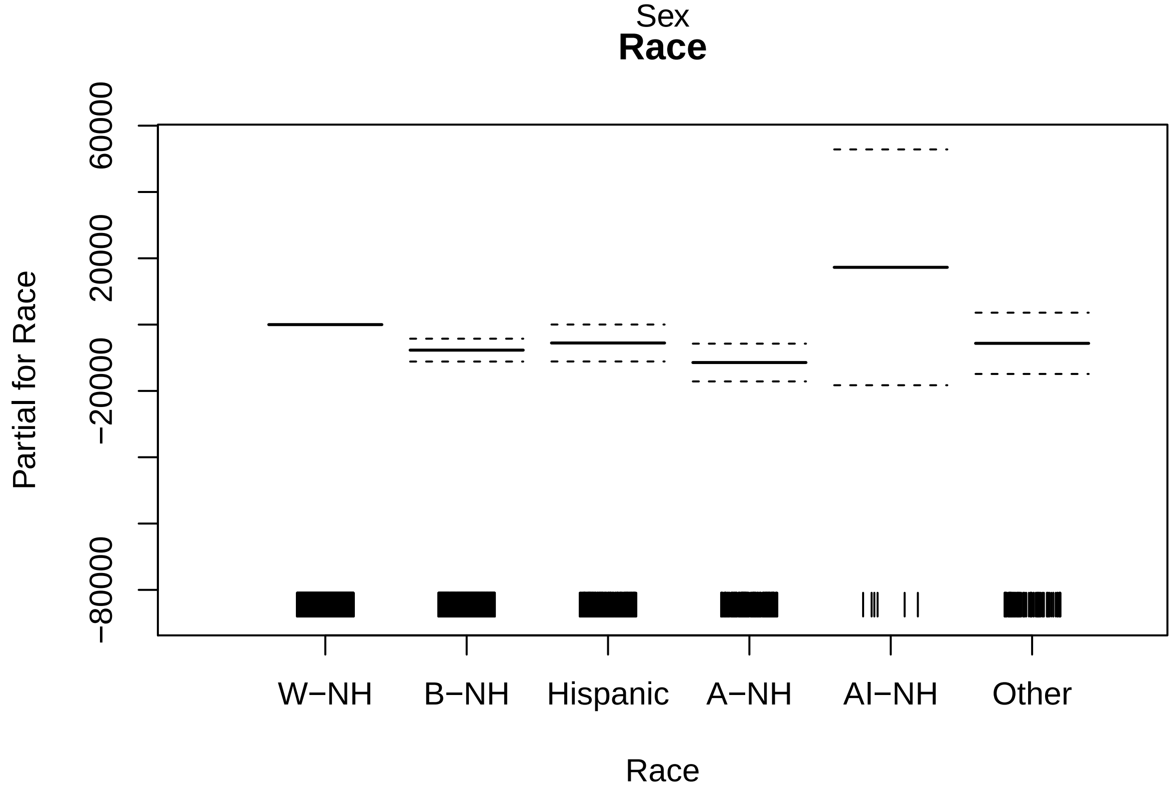
Plot

```r
plot(fit_gam2, all.terms = TRUE)
```


s(Education,3.78)
−20000
0
20000
60000
0
5
10
15
20
Education


s(Age,3.12)
−20000
0
20000
60000
20
30
40
50
60
Age


Sex
Partial for Sex
60000
20000
−20000
−80000
Male
Female
Sex


Race
Partial for Race
60000
20000
−20000
−80000
W−NH
B−NH
Hispanic
A−NH
AI−NH
Other
Race

Logistic Regression

Estimate regression models predicting a categorical outcome

```
# Recode "Managers" to create a binary dependent variable
  acs <- acs  %>% mutate(manager = (as.numeric(fct_collapse(Managers,
                              yes = c("Chief Executives", "Managers", "Supervisors"),
                              no = "Other"))))

acs$manager[acs$manager == 2] <- 0
```

GLM

View summary

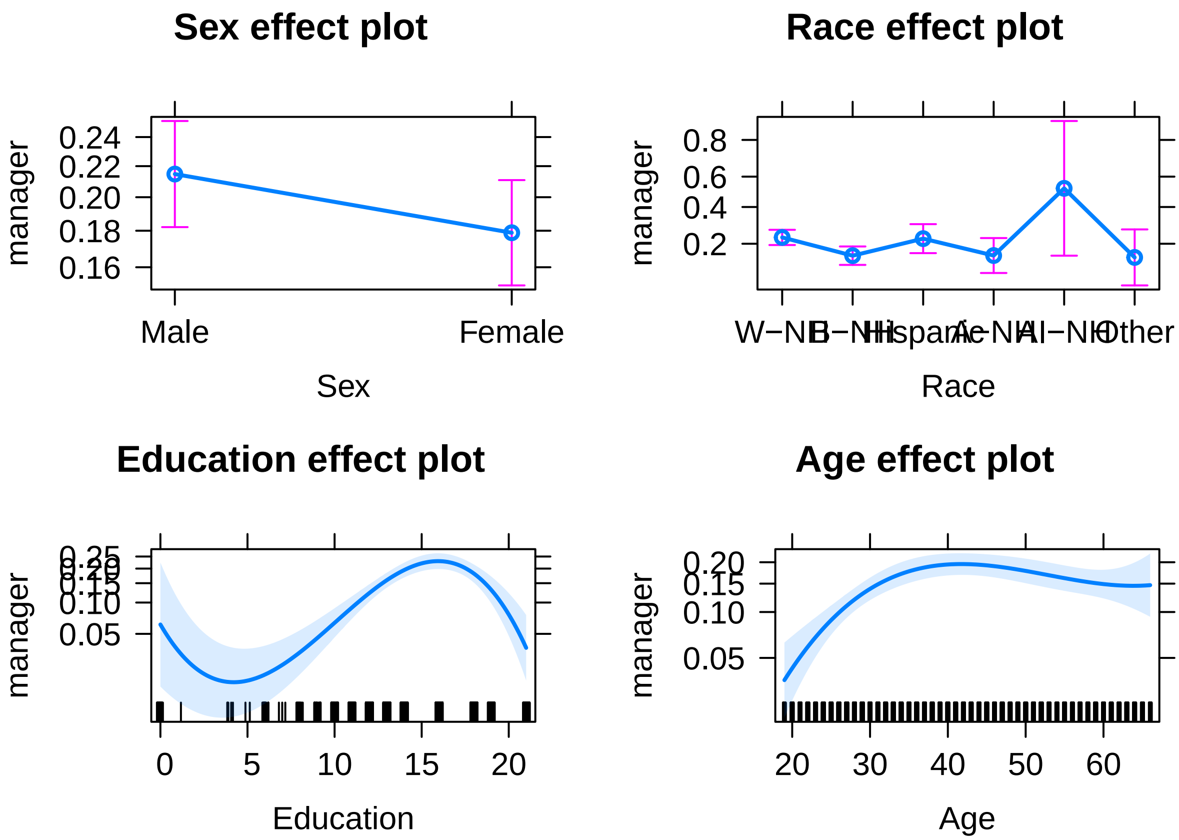
```
fit7 <- glm(manager ~ Sex + Race + bs(Education,3) + bs(Age,3),
          family = binomial(), data = acs)
summary(fit7)
```

```
##
## Call:
## glm(formula = manager ~ Sex + Race + bs(Education, 3) + bs(Age,
##     3), family = binomial(), data = acs)
##
## Deviance Residuals:
```

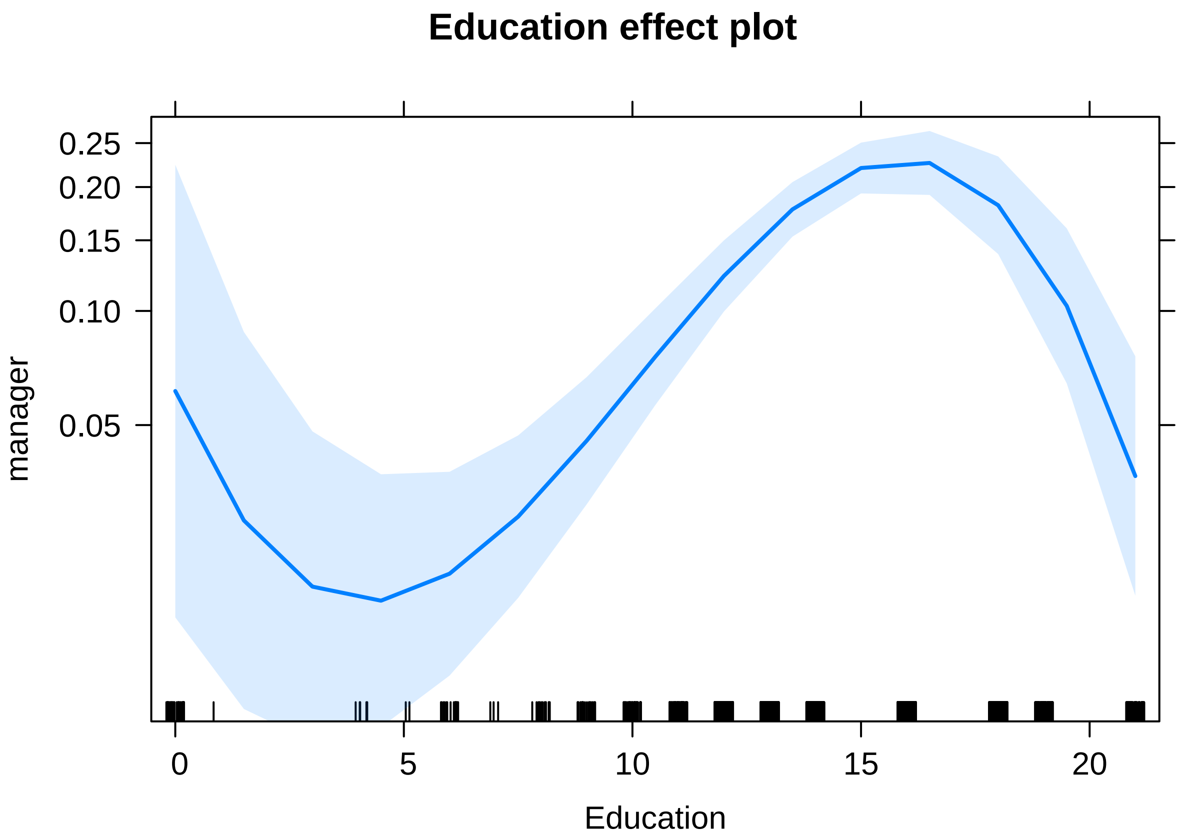
```
##      Min       1Q   Median       3Q      Max
## -1.0978  -0.5676  -0.4644  -0.3258   2.9181
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -4.28138    0.82616  -5.182 2.19e-07 ***
## SexFemale         -0.22792    0.10485  -2.174 0.029722 *
## RaceB-NH          -0.48750    0.12939  -3.768 0.000165 ***
## RaceHispanic      -0.03247    0.20003  -0.162 0.871044
## RaceA-NH          -0.48329    0.23257  -2.078 0.037700 *
## RaceAI-NH          1.31085    0.91376   1.435 0.151411
## RaceOther         -0.53398    0.38105  -1.401 0.161115
## bs(Education, 3)1 -5.02948    1.45817  -3.449 0.000562 ***
## bs(Education, 3)2  5.81360    1.03544   5.615 1.97e-08 ***
## bs(Education, 3)3 -0.55644    0.87548  -0.636 0.525048
## bs(Age, 3)1        3.13481    0.79886   3.924 8.71e-05 ***
## bs(Age, 3)2        1.38572    0.40058   3.459 0.000542 ***
## bs(Age, 3)3        1.54642    0.46583   3.320 0.000901 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2677.0  on 3597  degrees of freedom
## Residual deviance: 2543.9  on 3585  degrees of freedom
## AIC: 2569.9
##
## Number of Fisher Scoring iterations: 5
```

Plot

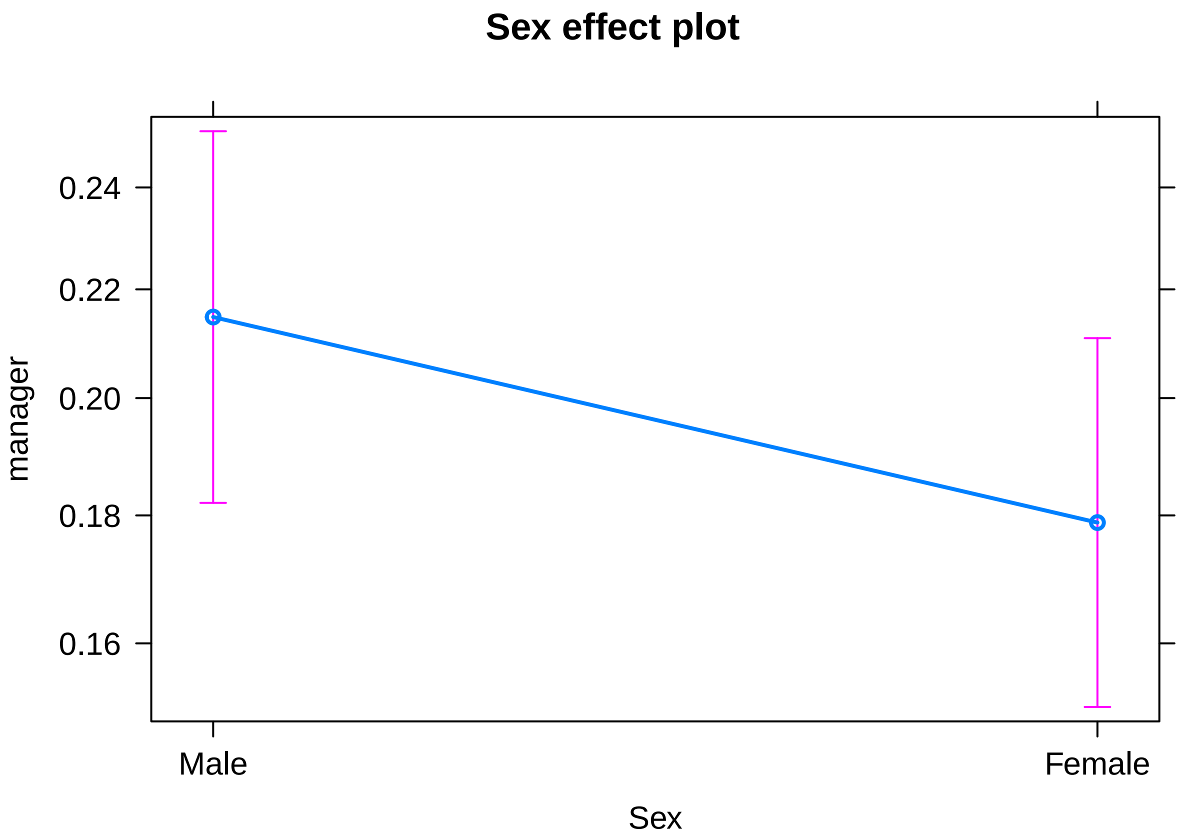
```r
plot(allEffects(fit7, xlevels = 50))
```



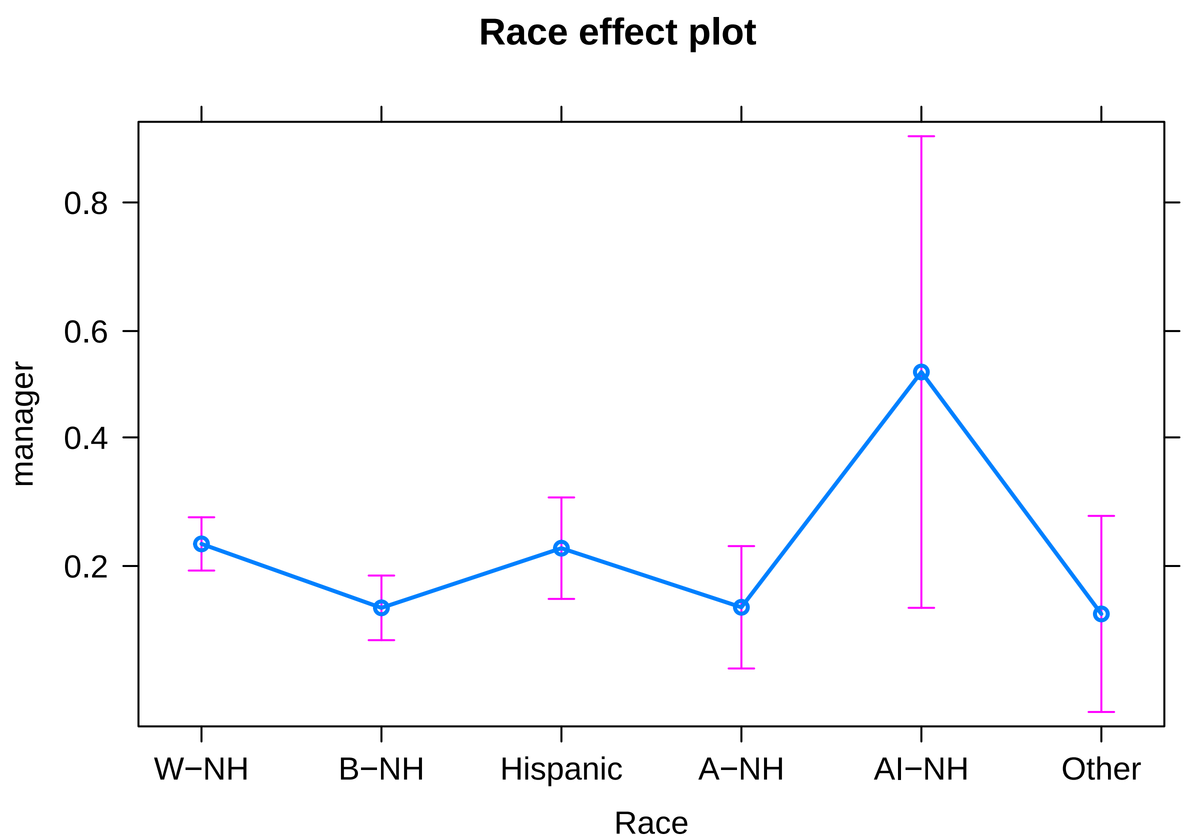

```r
plot(effect("bs(Education,3)", fit7))
```




```r
plot(effect("Sex", fit7))
```




```
plot(effect("Race", fit7))
```



GAM

View summary

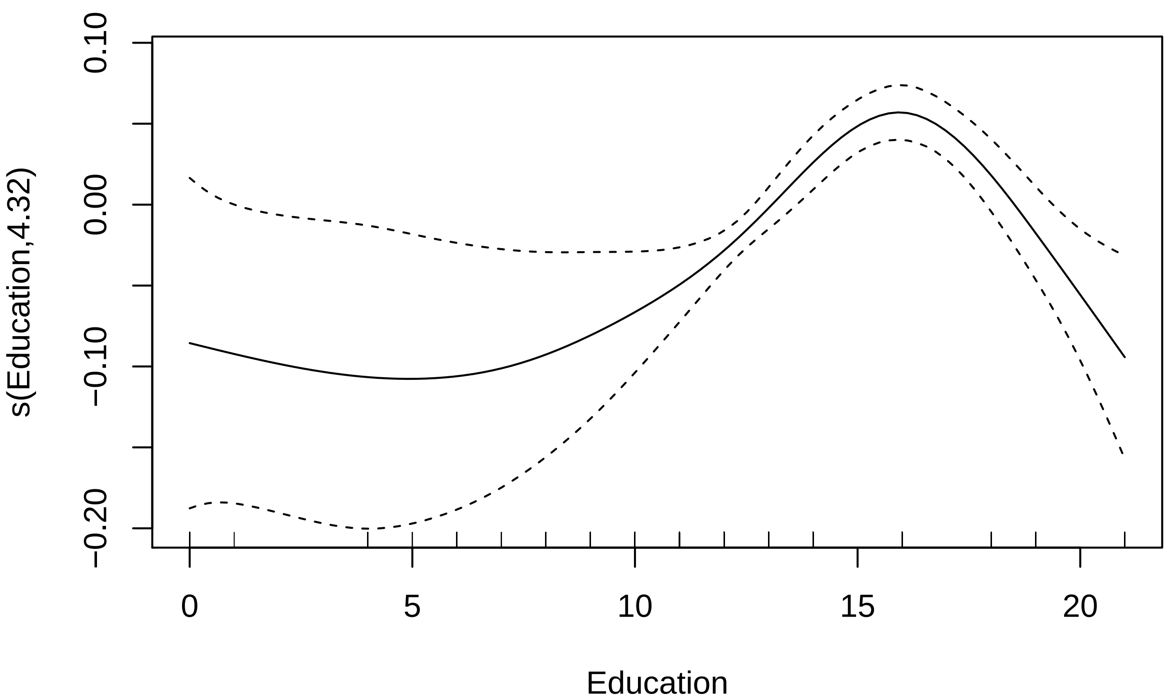
```r
fit_gam3 <- gam(manager ~ Sex + Race + s(Education) + s(Age), data = acs)
summary(fit_gam3)
```

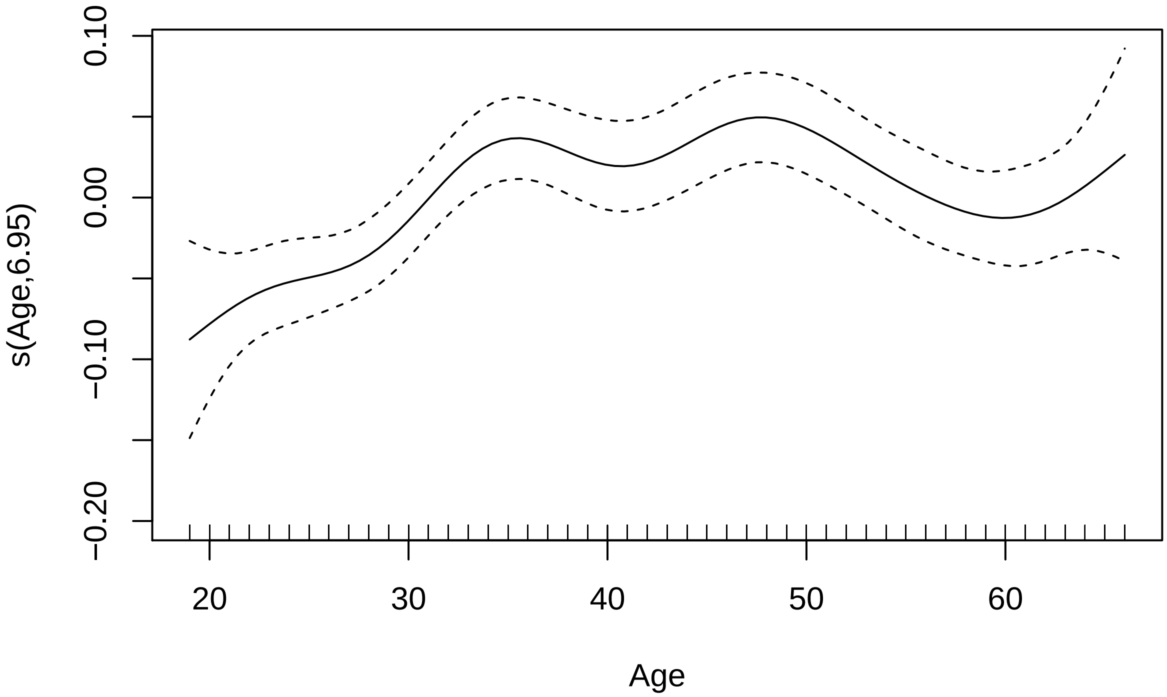
```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## manager ~ Sex + Race + s(Education) + s(Age)
##
## Parametric coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.155474   0.009441  16.469  < 2e-16 ***
## SexFemale    -0.023123   0.010947  -2.112 0.034736 *
## RaceB-NH     -0.048331   0.012897  -3.748 0.000181 ***
## RaceHispanic -0.007726   0.020890  -0.370 0.711533
## RaceA-NH     -0.048312   0.021349  -2.263 0.023699 *
## RaceAI-NH     0.209678   0.133225   1.574 0.115607
## RaceOther    -0.053692   0.034632  -1.550 0.121142
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

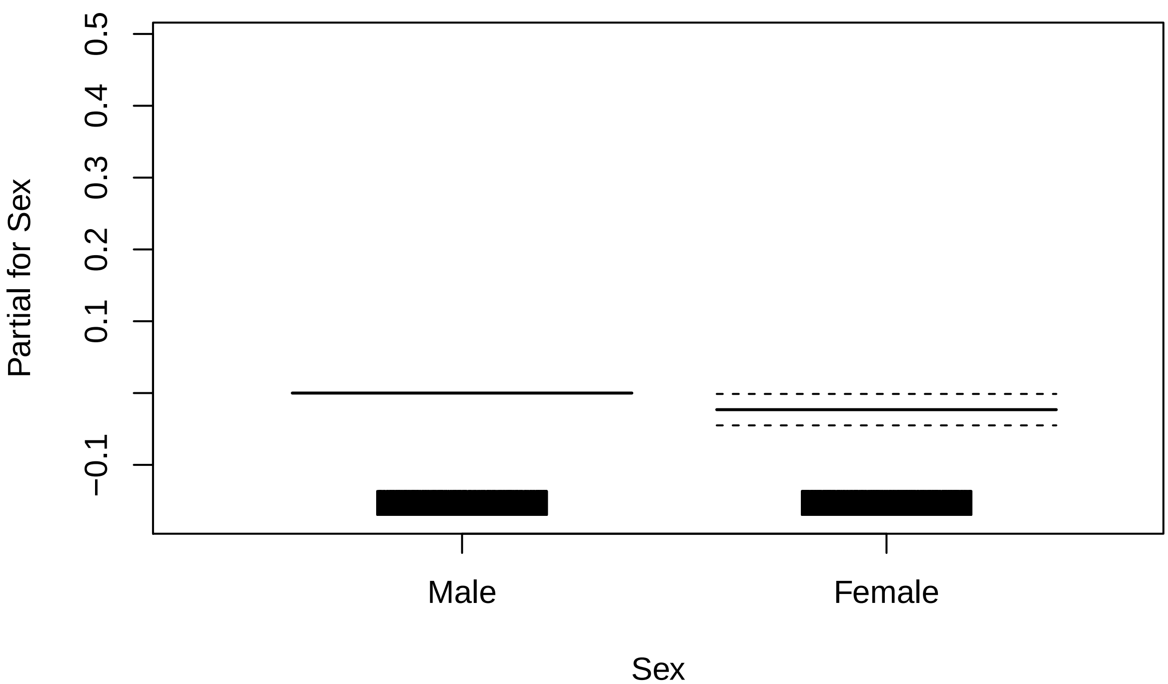
```
##
## Approximate significance of smooth terms:
##                 edf Ref.df      F  p-value
## s(Education) 4.323  5.233 10.404 3.80e-10 ***
## s(Age)       6.952  8.033  5.482 5.79e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.0328   Deviance explained = 3.75%
## GCV = 0.10457  Scale est. = 0.10404   n = 3598
```

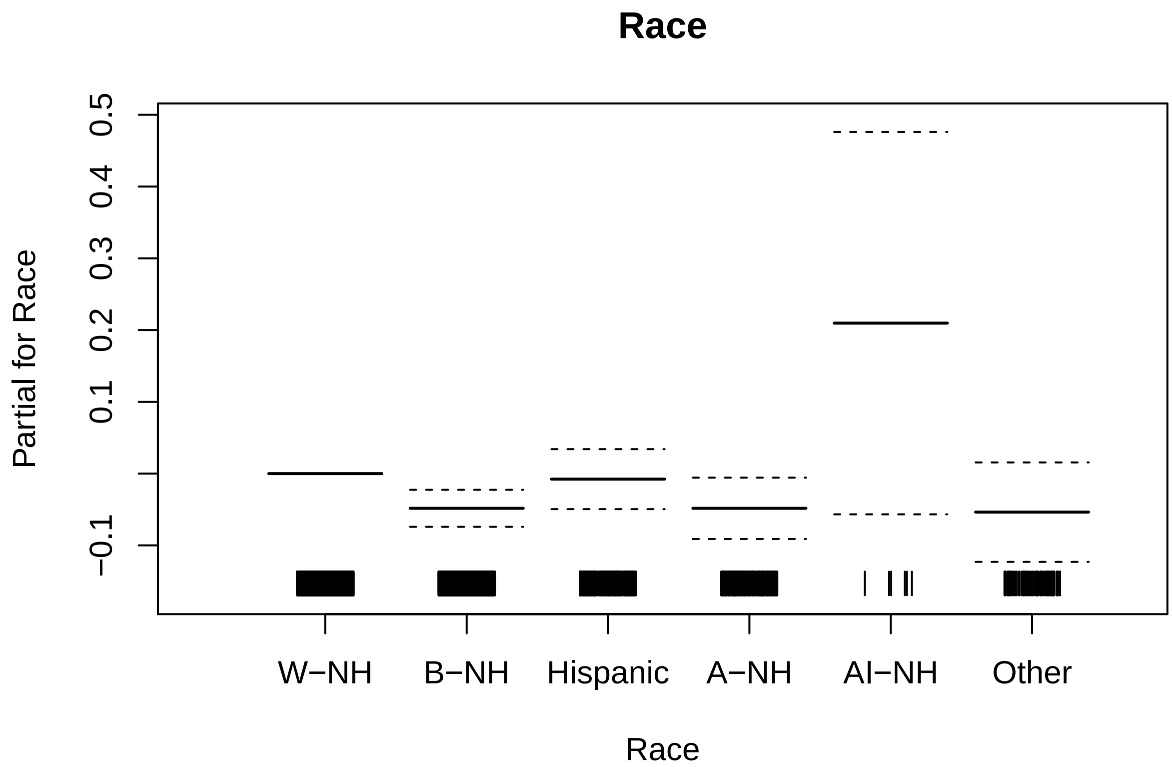
Plot

```r
plot(fit_gam3, all.terms = TRUE)
```




0.10
0.00
−0.10
−0.20
s(Age,6.95)
20
30
40
50
60
Age


Sex
0.5
0.4
0.3
0.2
0.1
−0.1
Partial for Sex
Male
Female
Sex


Race
Partial for Race
0.5
0.4
0.3
0.2
0.1
−0.1
W−NH
B−NH
Hispanic
A−NH
AI−NH
Other
Race