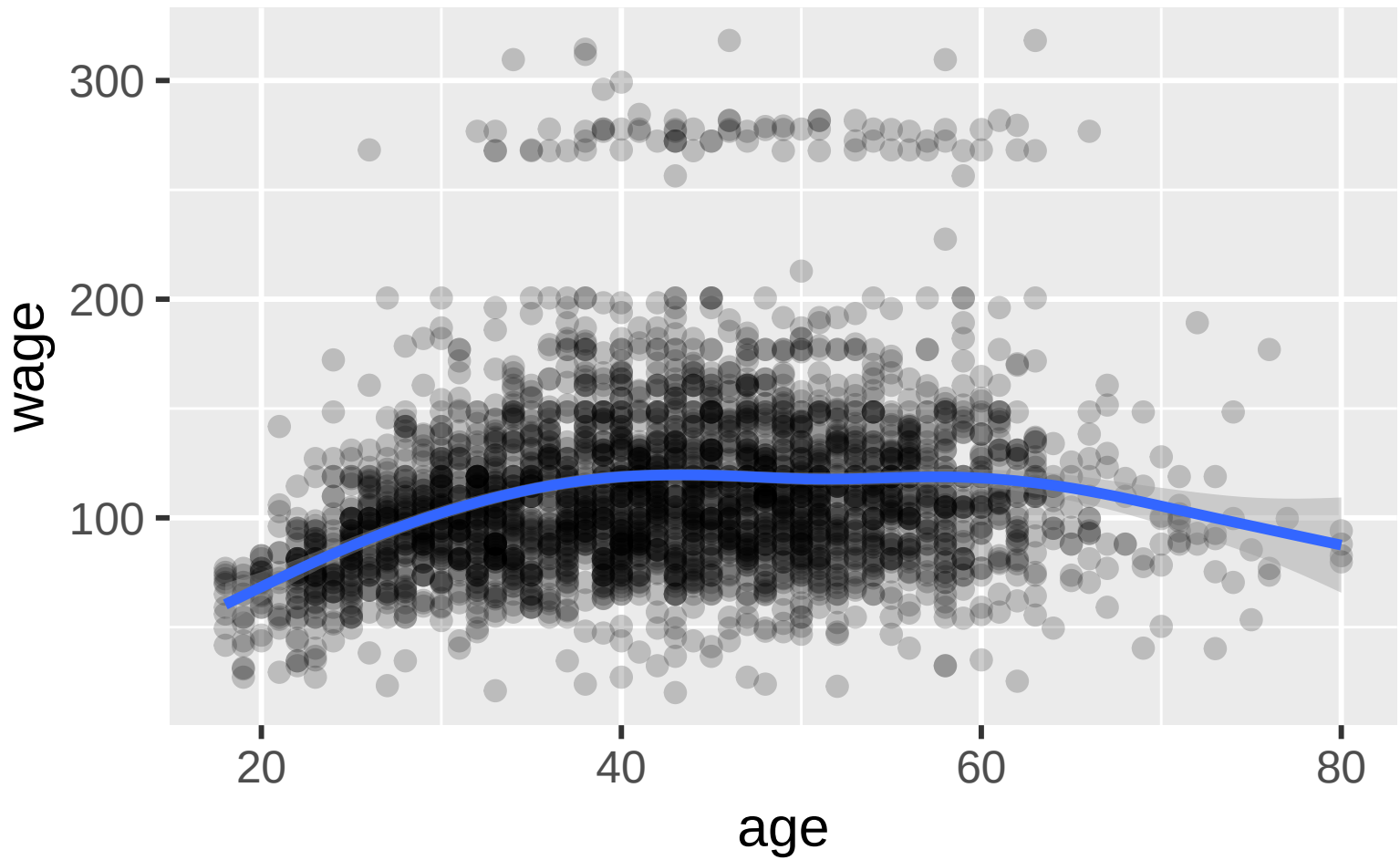
Statistics 360: Advanced R for Data Science

Multivariate Adaptive Regression Splines (MARS)

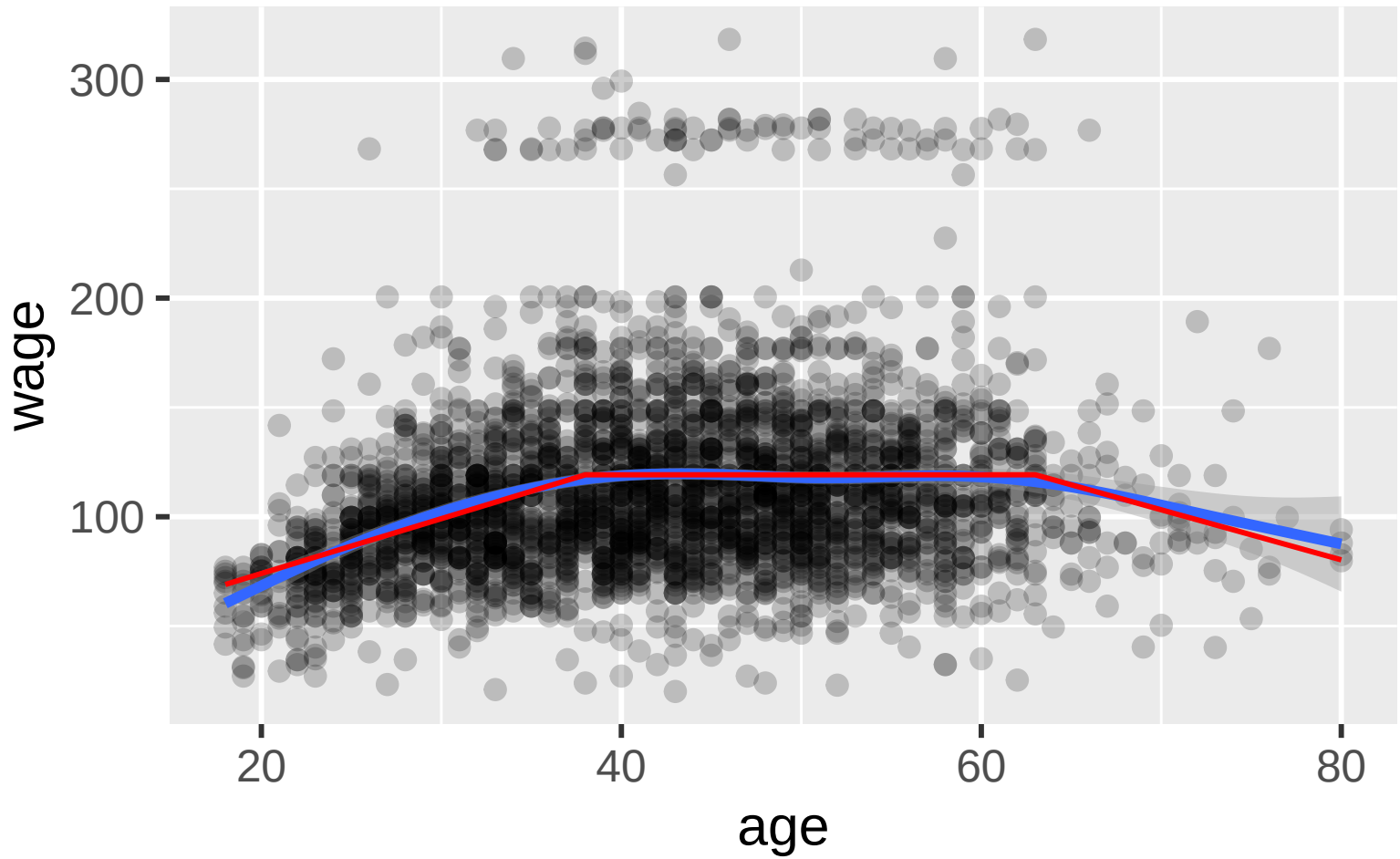
Brad McNeney

Example Data

```
library(tidyverse)
library(ISLR)
data(Wage) # help(Wage) for info
ggplot(Wage,aes(x=age,y=wage)) + geom_point(alpha=.2) + geom_smooth()
```



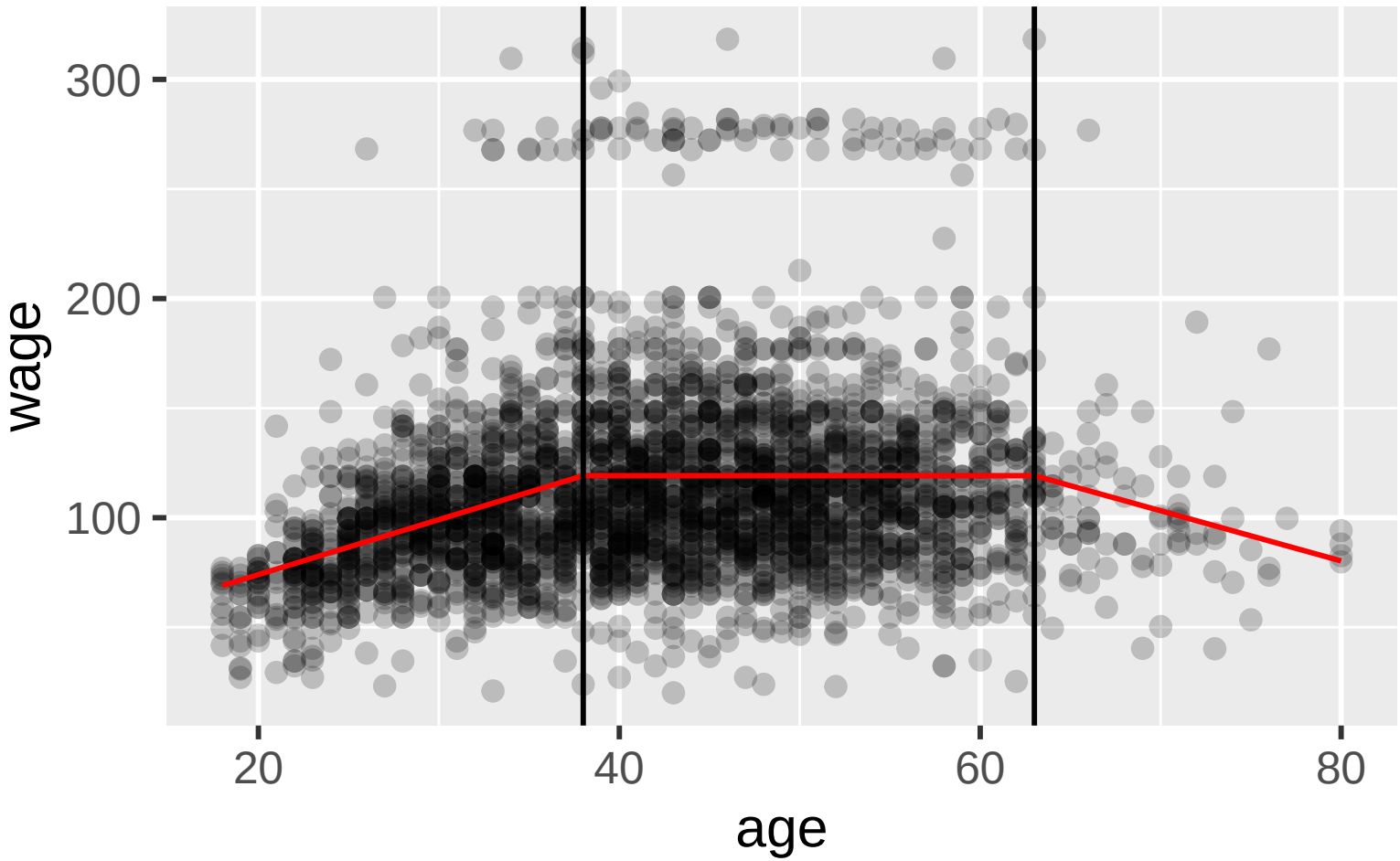
```
library(earth)
ee <- earth(wage ~ age, data=Wage)
Wage <- mutate(Wage,pwage = predict(ee))
ggplot(Wage,aes(x=age,y=wage)) + geom_point(alpha=.2) + geom_smooth()+
  geom_line(aes(y=pwage),color="red")
```



```
summary(ee)
```

```
## Call: earth(formula=wage~age, data=Wage)
##
##             coefficients
## (Intercept)   119.190151
## h(38-age)      -2.508377
## h(age-63)      -2.289070
##
## Selected 3 of 4 terms, and 1 of 1 predictors
## Termination condition: RSq changed by less than 0.001 at 4 terms
## Importance: age
## Number of terms at each degree of interaction: 1 2 (additive model)
## GCV 1595.44    RSS 4770379    GRSq 0.08405764    RSq 0.08649934
```

```
ggplot(Wage,aes(x=age,y=wage)) + geom_point(alpha=.2) +
  geom_line(aes(y=pwage),color="red") +
  geom_vline(xintercept=38) +
  geom_vline(xintercept=63)
```



Hinge functions

- The points 38 and 63 are "knots" where the piece-wise linear function changes slope.
- The piece-wise linear fit is a linear model in a constant term (intercept) and two "hinge" functions, $h(38 - age)$ and $h(age - 63)$, where

$$h(x) = \max(0, x)$$

- Hinge functions $h(x - c)$ and $h(c - x)$ are called mirror image.
  - Exercise: Plot two mirror-image hinge functions for `x <- seq(from=0,to=50,length=100)` and `c<-50`. Why are they called mirror image?

Fitting

- Once we are given the knots and hinge functions, the fit can be obtained by least squares.

```
Wage <- mutate(Wage,h1=pmax(0,38-age),h2=pmax(0,age-63))
ff <- lm(wage ~ h1+h2,data=Wage)
summary(ff)$coefficients
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 119.190151  0.8539353 139.577500 0.000000e+00
## h1           -2.508377  0.1506008 -16.655805 1.216314e-59
## h2           -2.289070  0.5949343  -3.847601 1.217668e-04
```

Questions

- How do we choose the knots?
- What happens when there are multiple explanatory variables, and we allow for interactions between them?