Ch. 08 Introduction

This example uses the "Current Population Survey" (CPS) dataset. There are 61395 observations.

```
> cps =
read.csv("http://home.cc.umanitoba.ca/~godwinrt/3180/data/cps.csv")
```

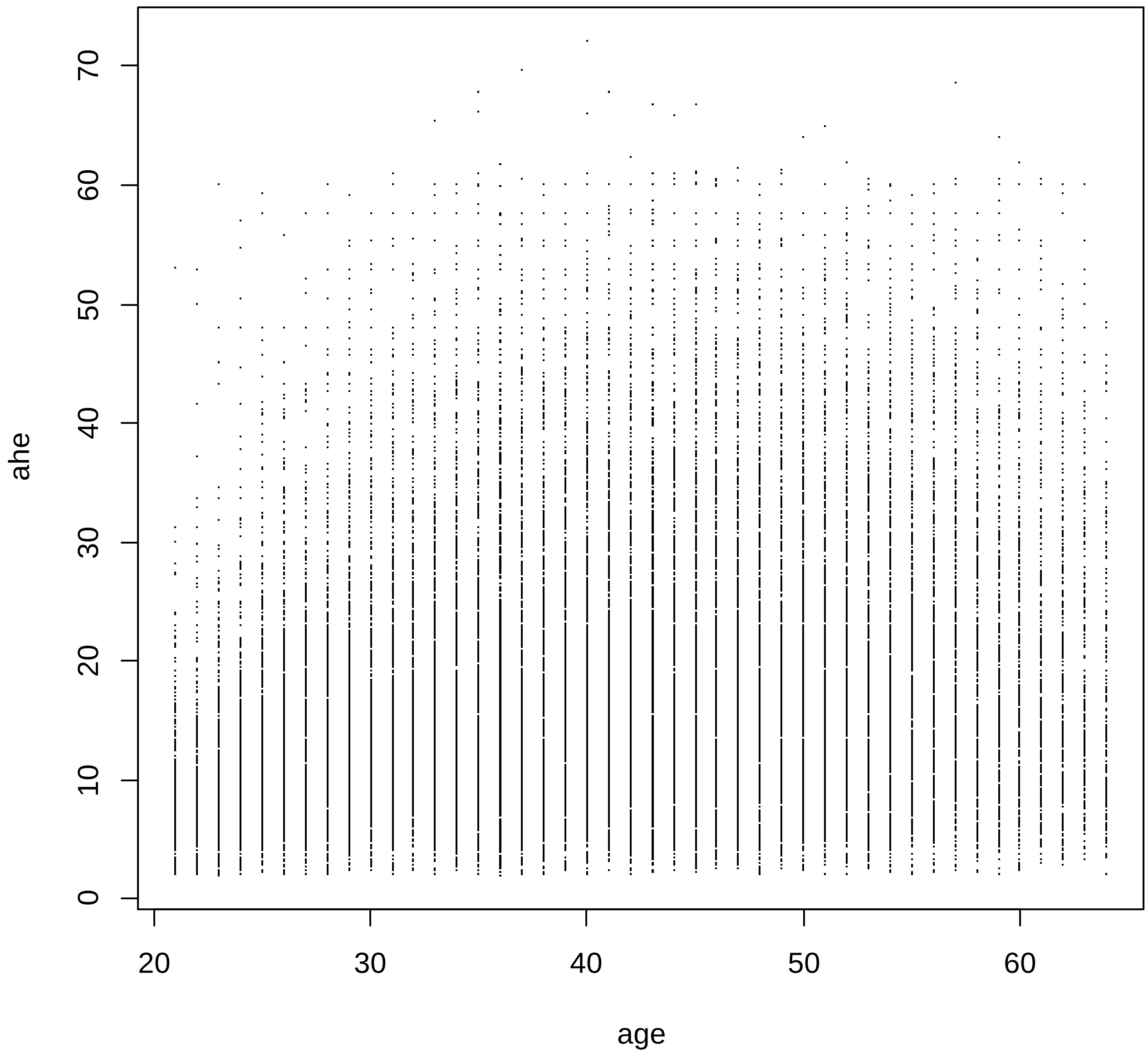
```
> attach(cps)
```

```
> head(cps)
```

```
        ahe female age northeast midwest south west yrseduc
1 20.673077      0  31         0       0     1    0      14
2 24.278847      0  50         0       0     1    0      12
3 10.149572      0  36         0       0     1    0      12
4  8.894231      1  33         0       0     1    0      10
5  6.410256      1  56         0       0     1    0      10
6 16.666666      1  52         0       0     1    0      12
```

View the relationship between *age* and *ahe*:

```
plot(age, ahe, pch = ".")
```


ahe
0
10
20
30
40
50
60
70
age
20
30
40
50
60

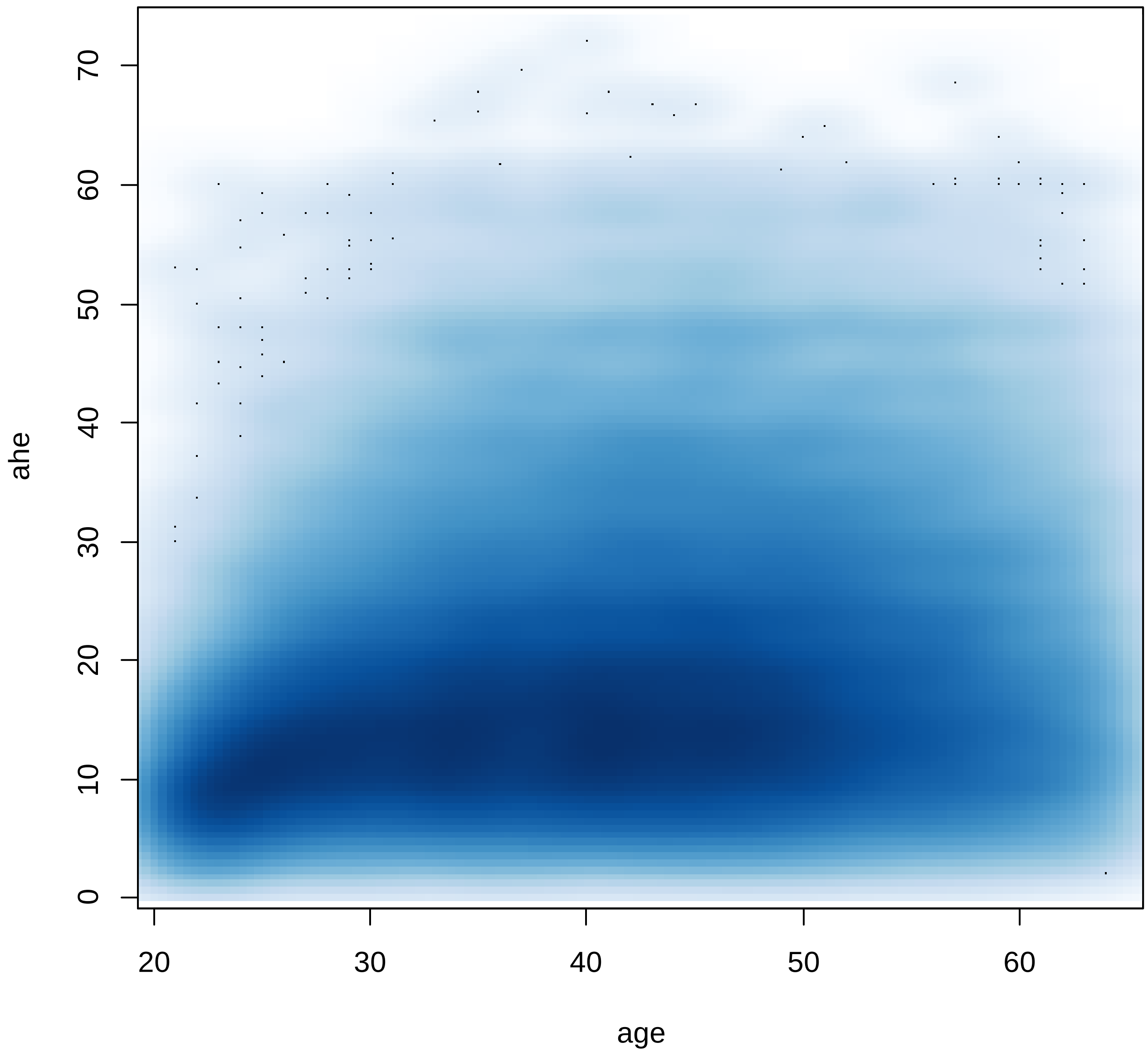
There are too many observations to see what's going on. A useful command for large datasets is:

```
smoothScatter(age,ahe)
```

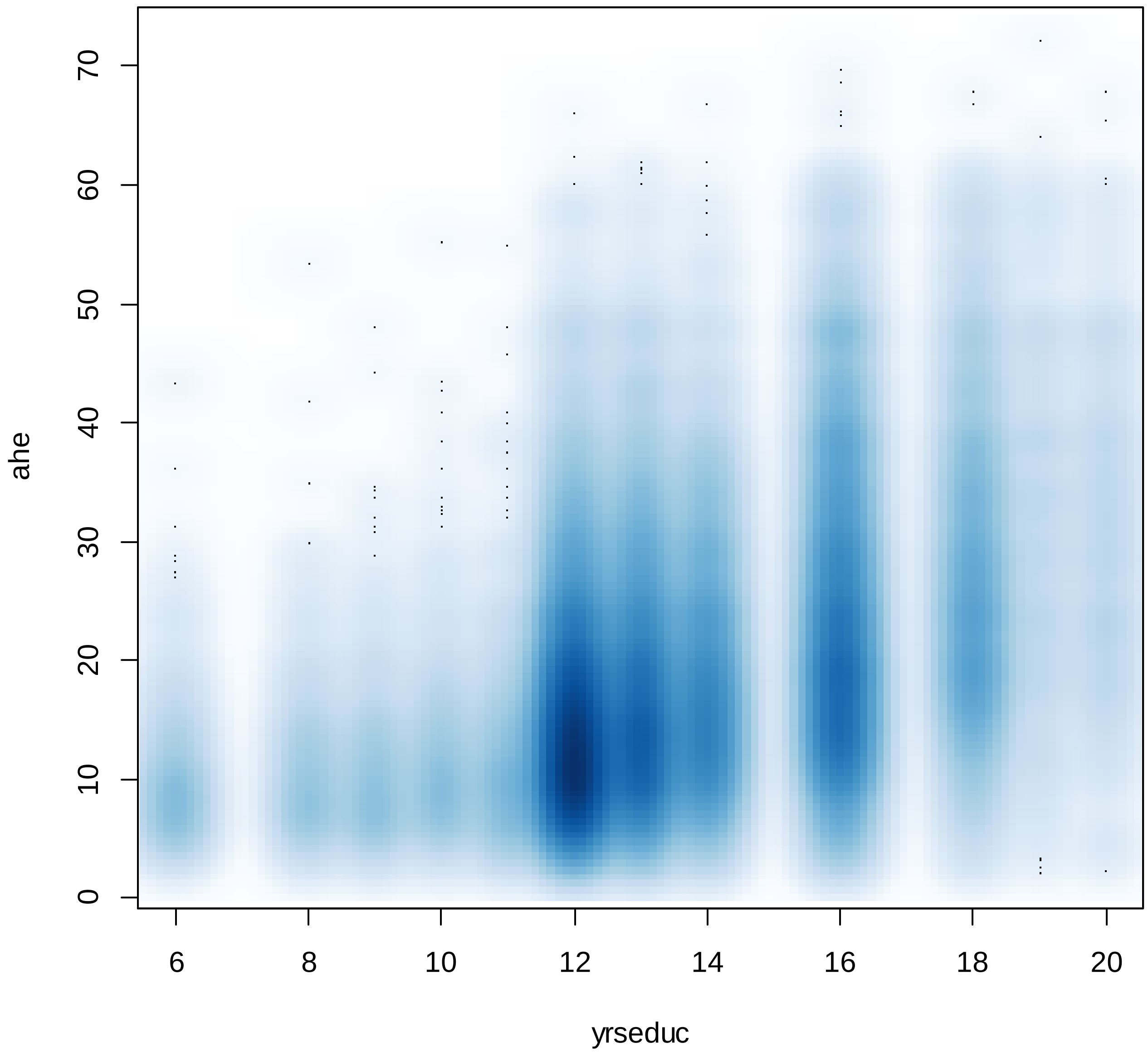
Let's also look at the relationship between ahe and yrseduc:

```
smoothScatter(yrseduc,ahe)
```

As you look at these relationships, imagine trying to fit a regression "line" through the data.



ahe
0
10
20
30
40
50
60
70
20
30
40
50
60
age



ahe
0
10
20
30
40
50
60
70
6
8
10
12
14
16
18
20
yrseduc

It appears that *age* and *yrseduc* have a non-linear effect on *ahe*. In fact, many effects in economics are non-linear. For example, diminishing marginal utility, and increasing / decreasing returns-to-scale.

In such cases, the effect on Y of a change in X depends on the value of X – that is, the marginal effect of X is not constant.

How can we capture this using our linear regression model? One idea is based on a Taylor series approximation. See:
http://en.wikipedia.org/wiki/Taylor_series#/media/File:Exp_series.gif

We won’t discuss the Taylor series here.

The idea is that non-linear functions can be **approximated** using **polynomials**. For example, a polynomial function is:

$$y = a + bx + cx^2 + dx^3 + ex^4$$

This is a fourth-order polynomial. A second order polynomial is the familiar quadratic equation:

$$y = a + bx + cx^2$$

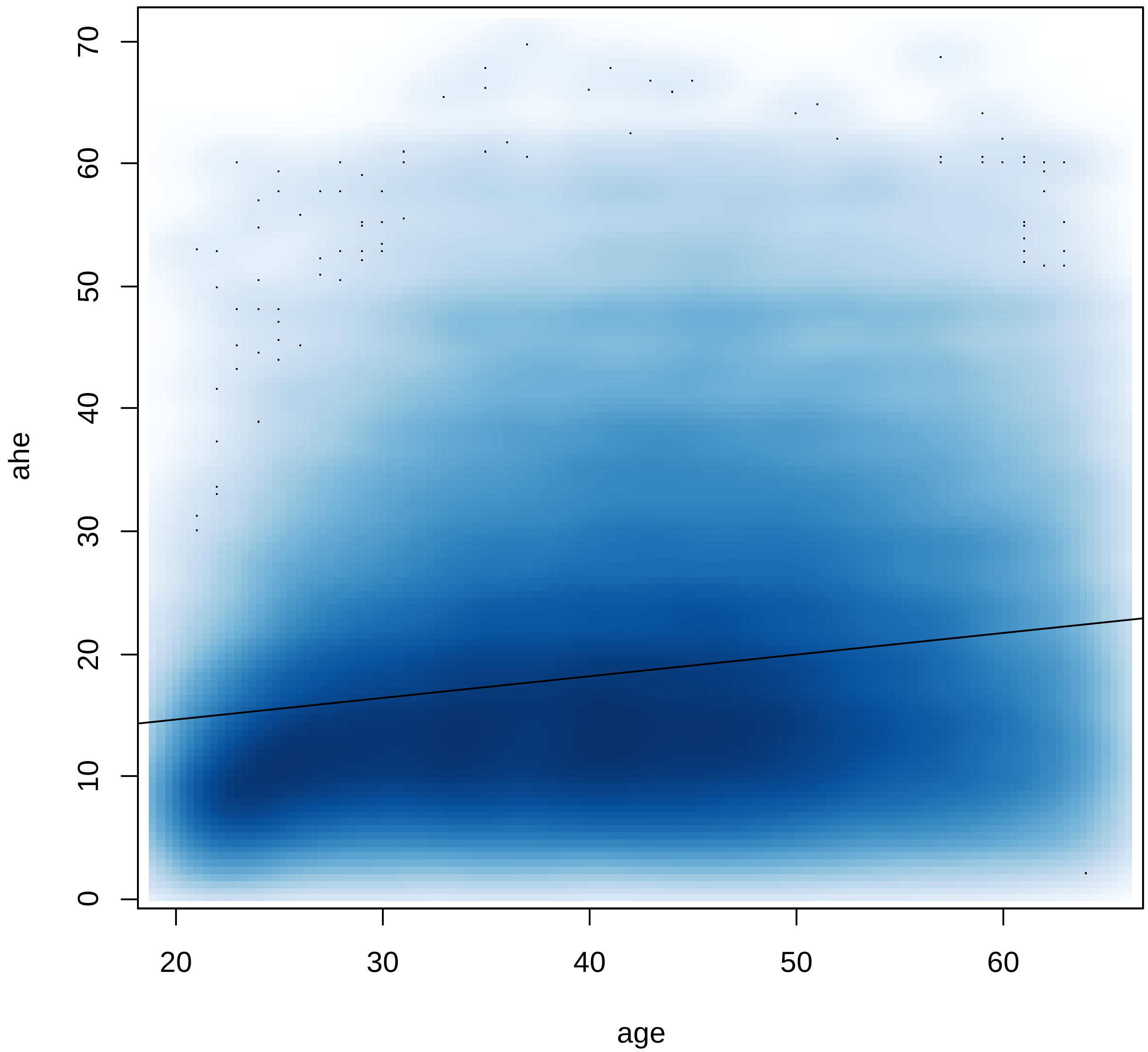
Now, let's try to capture the non-linear effect that *age* is having on *ahe*.

But first, let's see what happens when we fit a linear model:

```
par(new=T)

abline(lm(ahe ~ age))
```

It doesn't fit very well!


ahe
70
60
50
40
30
20
10
0
20
30
40
50
60
age

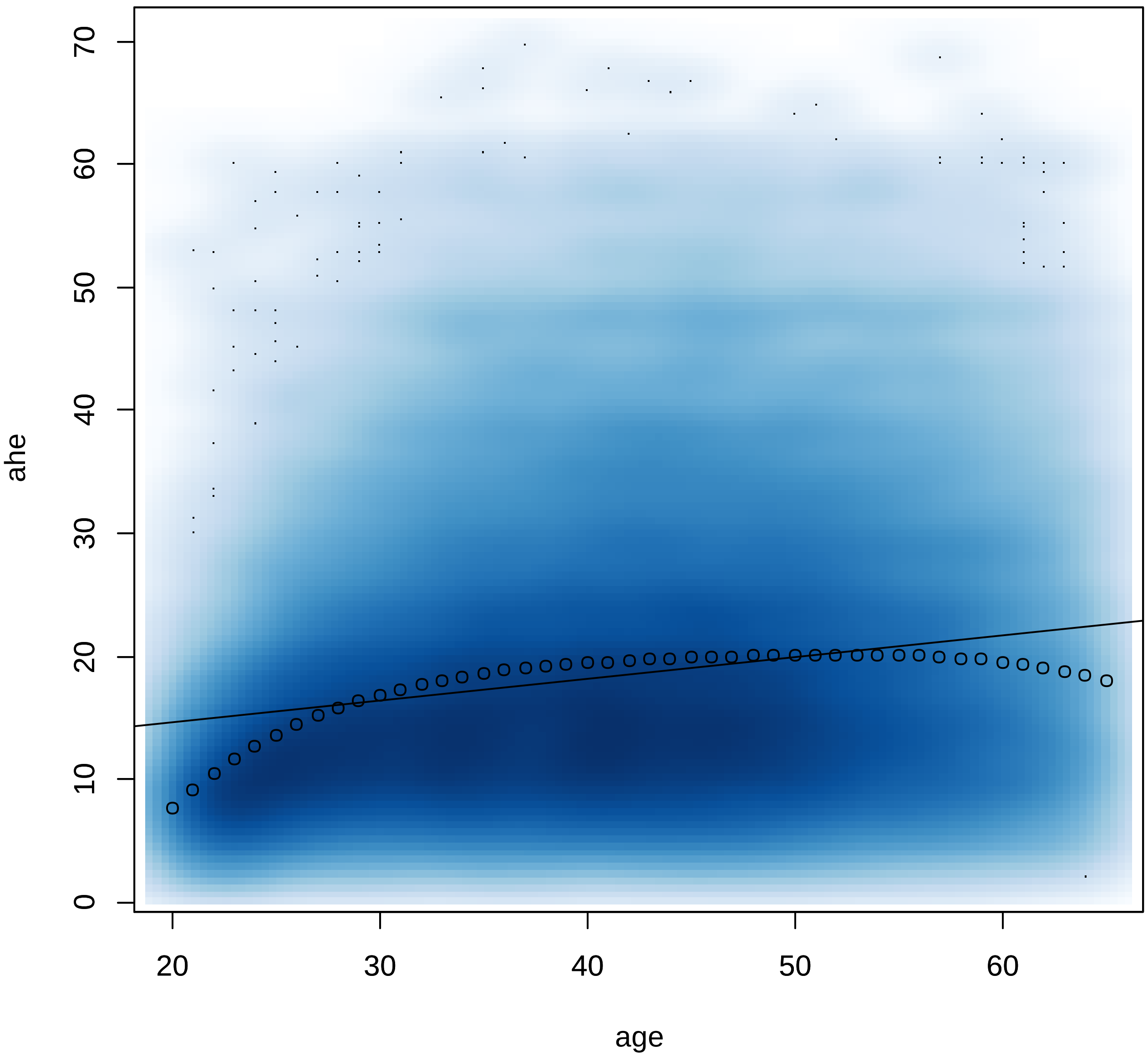
For the non-linear model, we first create new variables from *age*:

```
age2 = age^2
age3 = age^3
age4 = age^4
```

```
> summary(lm(ahe ~ age + age2 + age3 + age4))

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -6.905e+01  7.034e+00  -9.817  < 2e-16 ***
age          7.146e+00  7.371e-01   9.694  < 2e-16 ***
age2        -2.206e-01  2.795e-02  -7.892 3.01e-15 ***
age3         3.092e-03  4.559e-04   6.782 1.19e-11 ***
age4        -1.650e-05  2.706e-06  -6.097 1.09e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.842 on 61390 degrees of freedom
Multiple R-squared:  0.05551,   Adjusted R-squared:  0.05545
F-statistic:   902 on 4 and 61390 DF,  p-value: < 2.2e-16
```


ahe
age
0
10
20
30
40
50
60
70
20
30
40
50
60