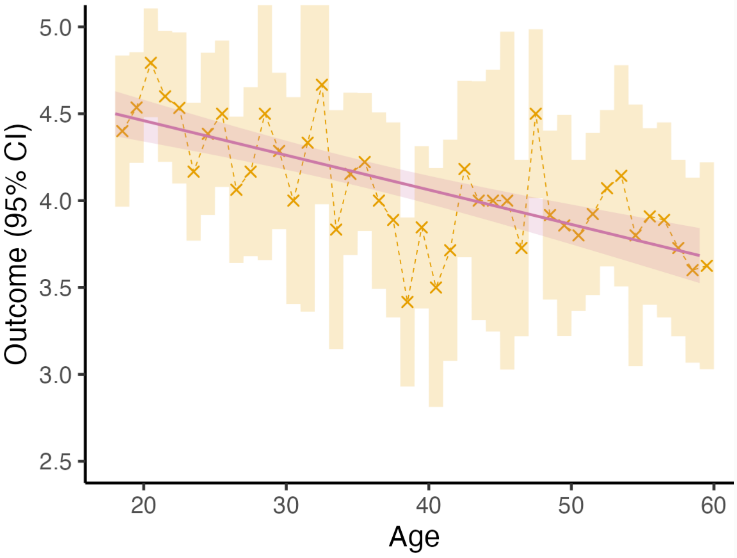
Different ways to model age



Let‘s say you want to include age as a covariate in a linear model – for example, because you consider it a confounder of some other association of interest. There are different ways to do this which vary in their flexibility and in the shape of the resulting trajectory.

One solution is to simply include age as a numerical predictor:

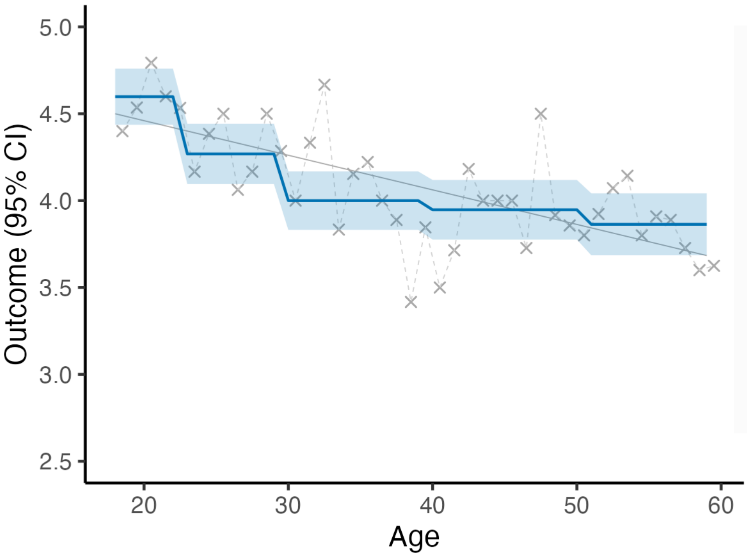
```
        lm(outcome ~ age)
```

This fits a straight line. The resulting uncertainty is quite low and including age costs 1 degree of freedom as a single parameter is additionally estimated – the slope of the straight line.

Another solution is to simply include age as a categorical predictor:

```
        lm(outcome ~ as.factor(age))
```

This estimates an individual value for each year of age. The resulting uncertainty is quite high and including age costs as many degrees of freedom as there are years of age, minus 1 (in this example, 41).

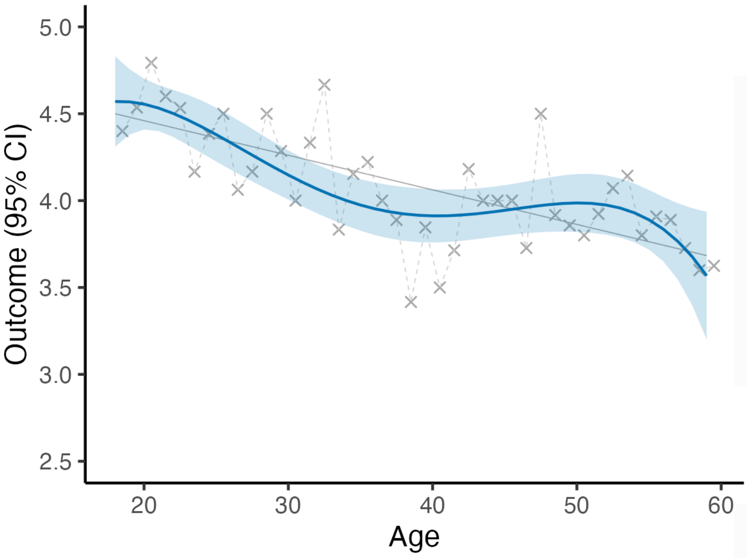


Other solutions lie somewhere in between these two extremes.

For example, we could form broader age categories:

```
        lm(outcome ~ I(age <= 22) +
                I(age > 22 & age <= 29) +
                I(age > 29 & age <= 39) +
                I(age > 39 & age <= 50))
```

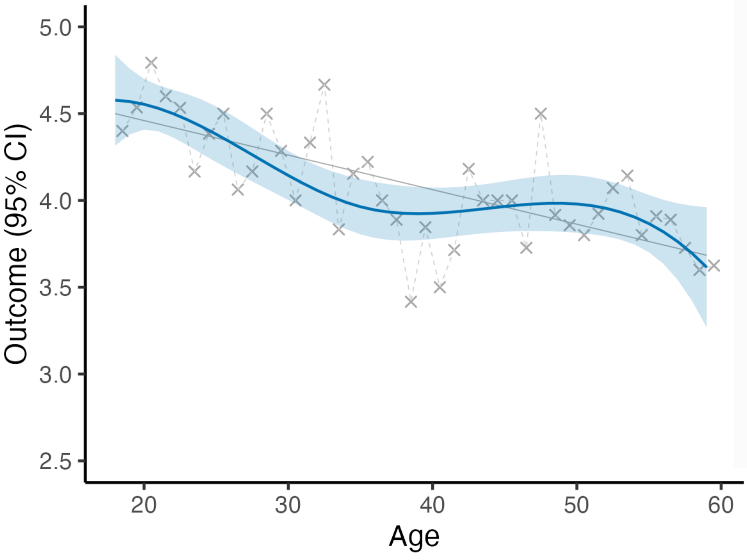
This estimates an individual value for each category. The resulting uncertainty is medium and including age costs as many degrees of freedom as there are categories, minus 1 (in this example, 5 – 1 = 4)



We could also use a polynomial:

```
        lm(outcome ~ I(age^4) +
                I(age^3) +
                I(age^2) +
                age)
```

This fits a fourth-degree polynomial. The resulting uncertainty is medium and including age costs as many degrees of freedom as the degree of the polynomial, so in this case 4.



And we could also use splines:

```
        library(splines)
        lm(outcome ~ bs(age, df = 4))
```

These result in a trajectory that is locally smoothed. Under the hood, multiple synthetic variables are calculated as tranformations of age and then included as predictors. The resulting uncertainty is medium and including age costs as many degrees of freedom as there are synthetic variables, in this case 4.

Even though splines and polynomials result in regression outputs that look very different, note how here they arrive at almost precisely the same age trajectory.