Stats 506, F20, Group Project

Group 6: Erin Susan Cikanek, Suppapat Korsurat, Kyle William Schulz

November 13, 2020

Contents

Introduction

Linear regression has become widely known as a backbone of modern statistics. Even as more complex, “black box”-style machine learning techniques increase in popularity, many statisticians and researchers still fall back on regression for its interpretability and simpleness. However, linear regression relies on a number on assumptions that may not always be true in practice, such as the constant, monotonic linearity of predictor variables in relation to the response. In this guide, we explore the use of splines to help model predictor variables that may have changing relationships across their domain. These techniques help us to match the predictive power seen in some more advanced machine learning algorithms while keeping the benefits gained by using regression. We show examples in three popular statistical modelling languages - python, R, and STATA.

Data

In this guide, we will be using the “wage” dataset from the R package ISLR. This dataset contains wages from 3,000 Mid-Atlantic workers, along with a select number of other personal demographics. Our goal is to examine the relationship between these demographics and the worker’s raw yearly wage.

Method

We will first calculate a simple linear regression as a baseline. We will then implement four different spline-like techniques on the “age” predictor variable: a step function, polynomial regression, basis spline, and natural spline. At each step, we will check for fit quality, noting any potential improvements along the way. We will conclude with a retrospective and summary of what we learned.

Core Analysis

Python

```
"Kyle's Code"
```

```
"When you want to show only code, but prevent this chunck to run."
```

```
## [1] "When you want this chunck to run, but don't want to show the code."
```

Stata

```
"Erin's Code"
```

```
"When you want to show only code, but prevent this chunck to run."
```

```
## [1] "When you want this chunck to run, but don't want to show the code."
```

R

The library `splines` is required for implementing splines by using R.

```
library(splines)
```

First, considering the linear regression.

```
model <- lm(wage ~ age + education + year, data = data)
summary(model)
```

```
##
## Call:
## lm(formula = wage ~ age + education + year, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -113.323  -19.521   -3.964   14.438  219.172
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -2.058e+03  6.493e+02  -3.169  0.00154 **
## age                          5.621e-01  5.714e-02   9.838  < 2e-16 ***
## education2. HS Grad          1.140e+01  2.476e+00   4.603 4.34e-06 ***
## education3. Some College     2.423e+01  2.606e+00   9.301  < 2e-16 ***
## education4. College Grad     3.974e+01  2.586e+00  15.367  < 2e-16 ***
## education5. Advanced Degree  6.485e+01  2.804e+00  23.128  < 2e-16 ***
## year                         1.056e+00  3.238e-01   3.262  0.00112 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.89 on 2993 degrees of freedom
## Multiple R-squared:  0.2619, Adjusted R-squared:  0.2604
## F-statistic:   177 on 6 and 2993 DF,  p-value: < 2.2e-16
```

The R^2 is 0.2619, which is pretty low. Consider the scatter plot between `Wage` and `Age`.

The scatter plot show that the relationship between these two variables are not linear. Hence, we will try various types of spline.

Step Function

Consider applying the step function on `Age`.

```
model_cut <- lm(wage ~ cut(age, 4) + education + year, data = data)
summary(model_cut)
```

```
##
## Call:
## lm(formula = wage ~ cut(age, 4) + education + year, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -120.260  -19.442   -3.744   14.441  214.958
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -2408.5219   641.1663  -3.756 0.000176 ***
```

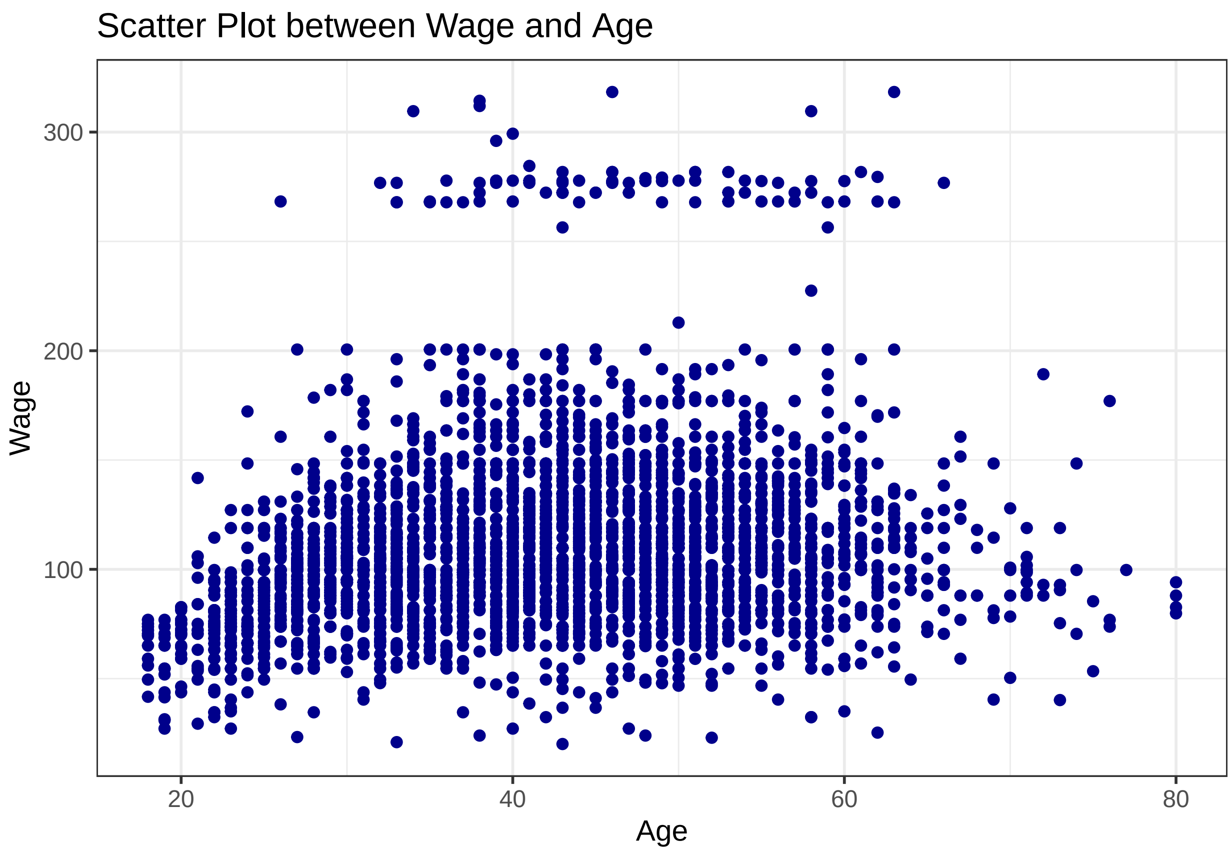


Figure 1: Figure 3.1 Scatter plot between Wage and Age

```
## cut(age, 4)(33.5,49]          20.9265     1.6085  13.010  < 2e-16 ***
## cut(age, 4)(49,64.5]          19.3732     1.8197  10.646  < 2e-16 ***
## cut(age, 4)(64.5,80.1]         8.0516     4.3783   1.839 0.066014 .
## education2. HS Grad           11.1534     2.4436   4.564 5.21e-06 ***
## education3. Some College      24.1620     2.5739   9.387  < 2e-16 ***
## education4. College Grad      39.2164     2.5533  15.359  < 2e-16 ***
## education5. Advanced Degree   64.1642     2.7675  23.185  < 2e-16 ***
## year                           1.2356     0.3197   3.865 0.000113 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.39 on 2991 degrees of freedom
## Multiple R-squared:  0.2828, Adjusted R-squared:  0.2809
## F-statistic: 147.4 on 8 and 2991 DF,  p-value: < 2.2e-16
```

The R^2 is 0.2828, which improved from the previous model. The plot below is a scatterplot between `Wage` and `Age`, also the yellow line represents the step function.

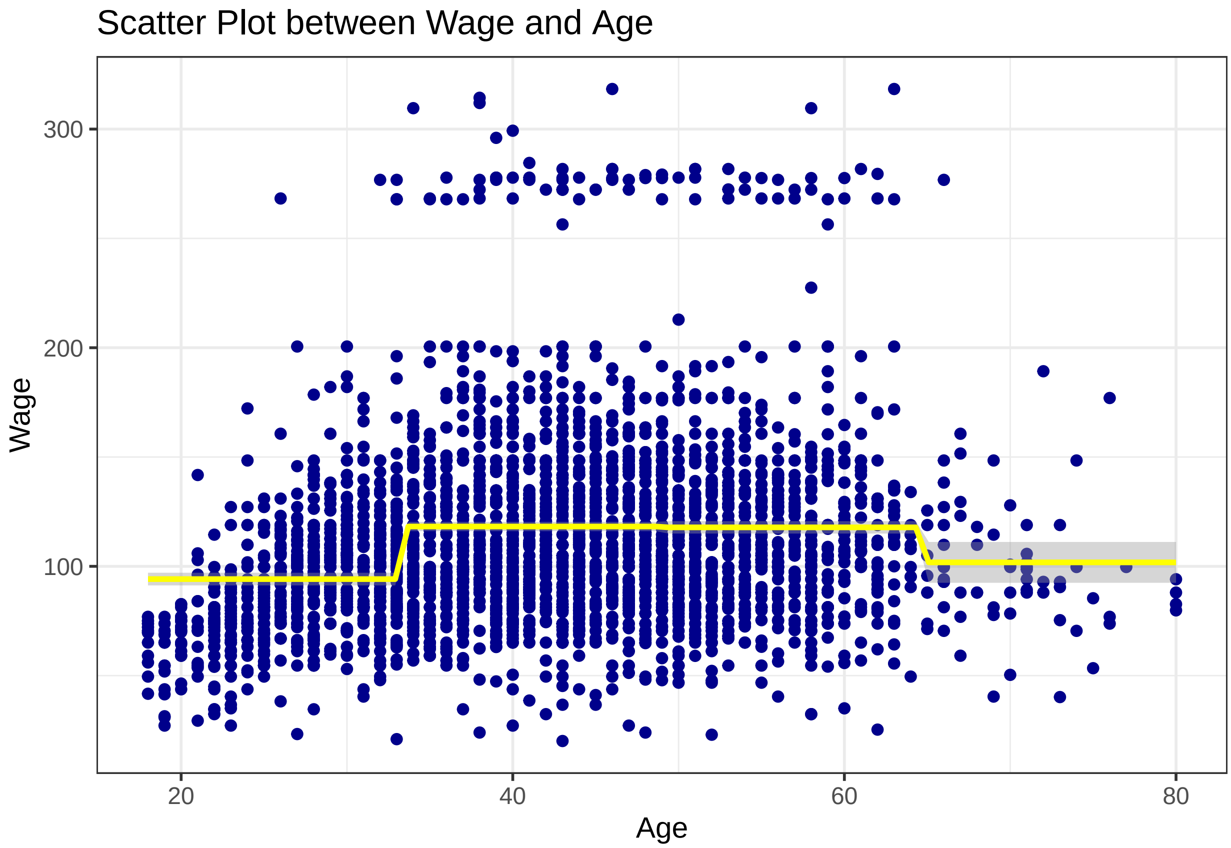
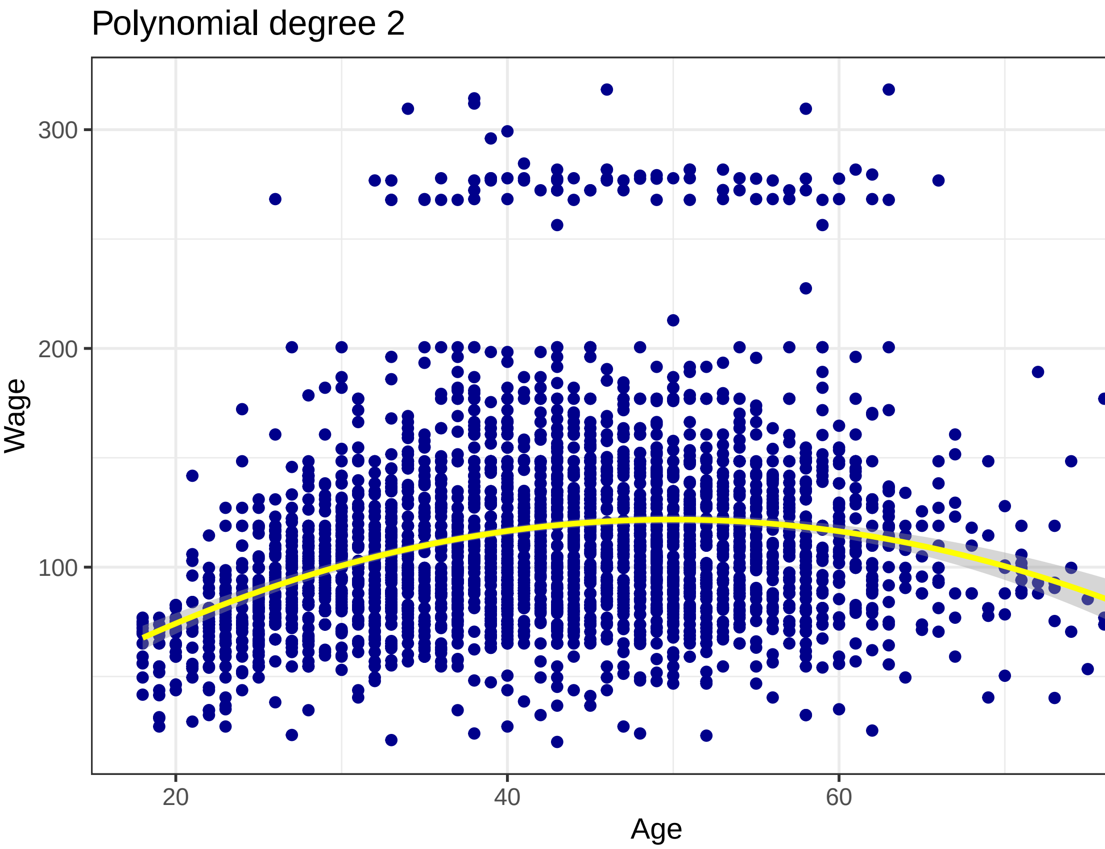


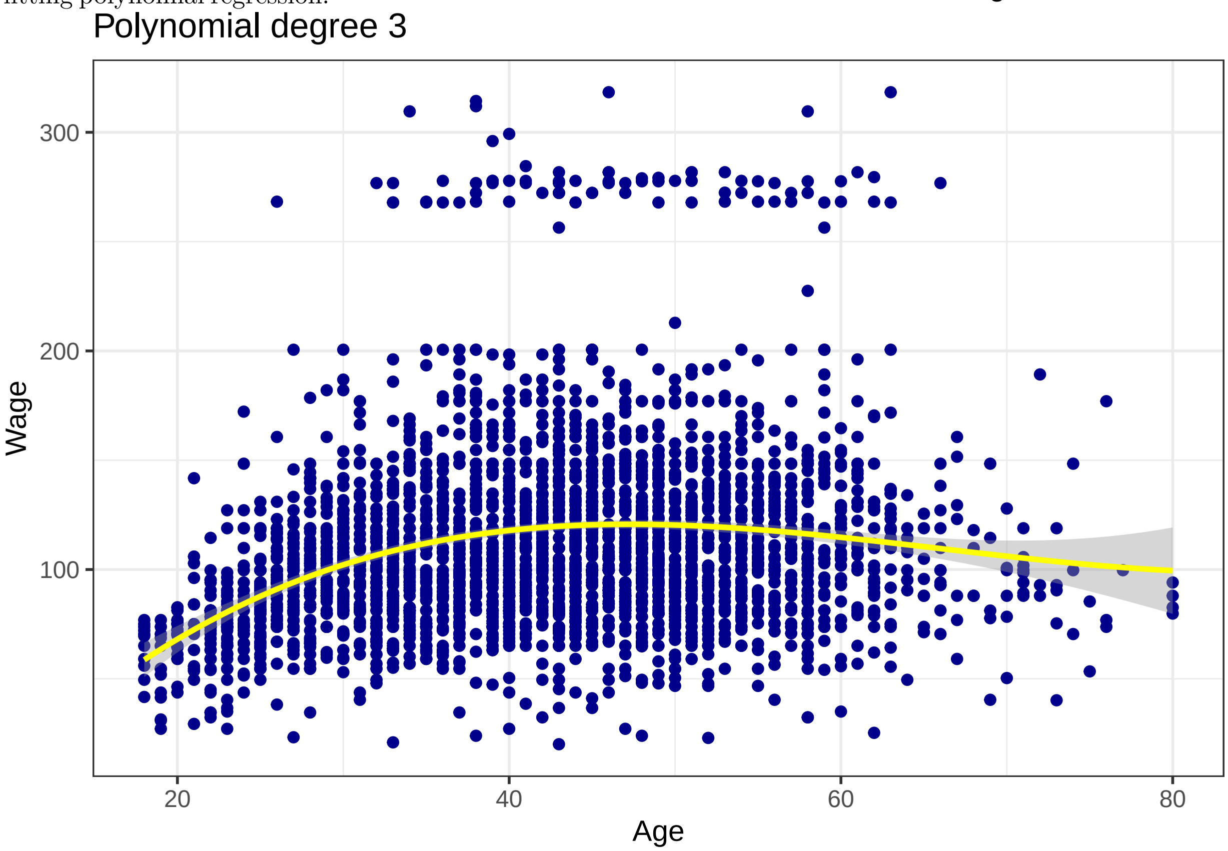
Figure 2: Figure 3.2 Scatter plot between Wage and Age with the step function.

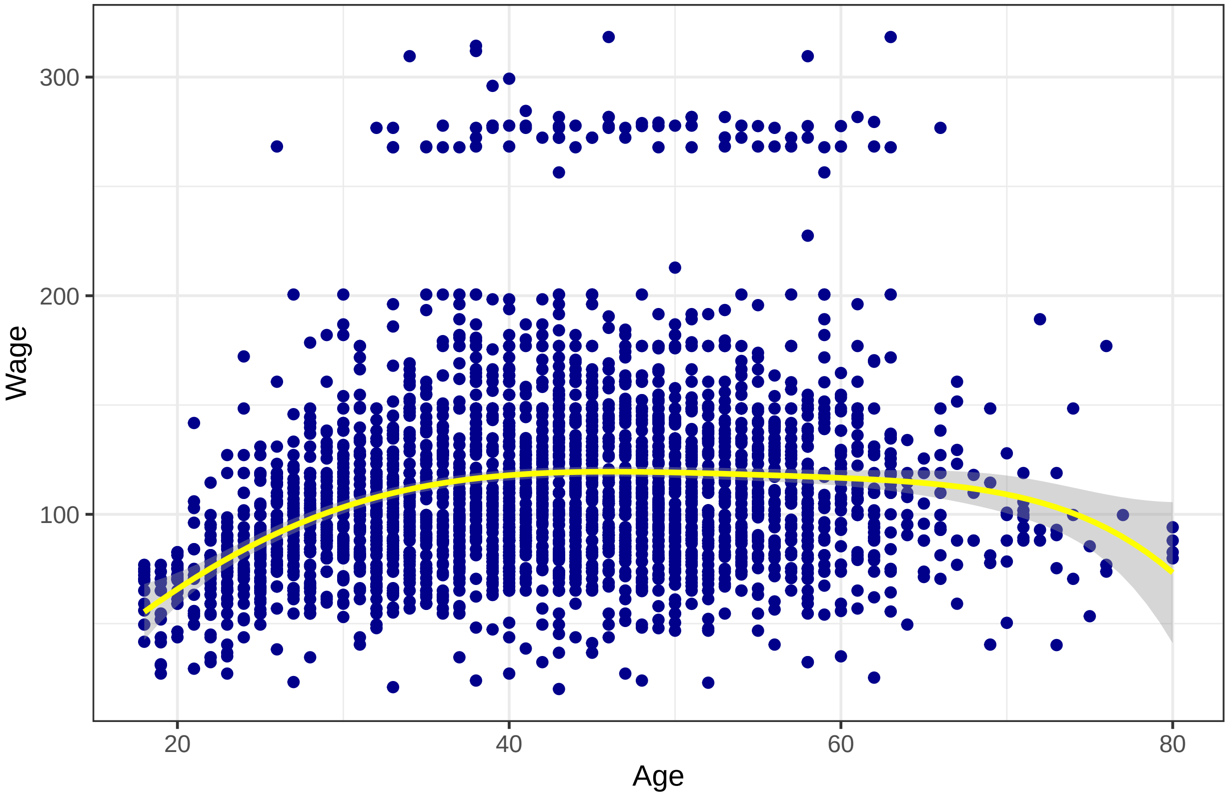
Polynomial Regression

Consider the various number for the degree in the polynomial regression. The plots below are the result from the

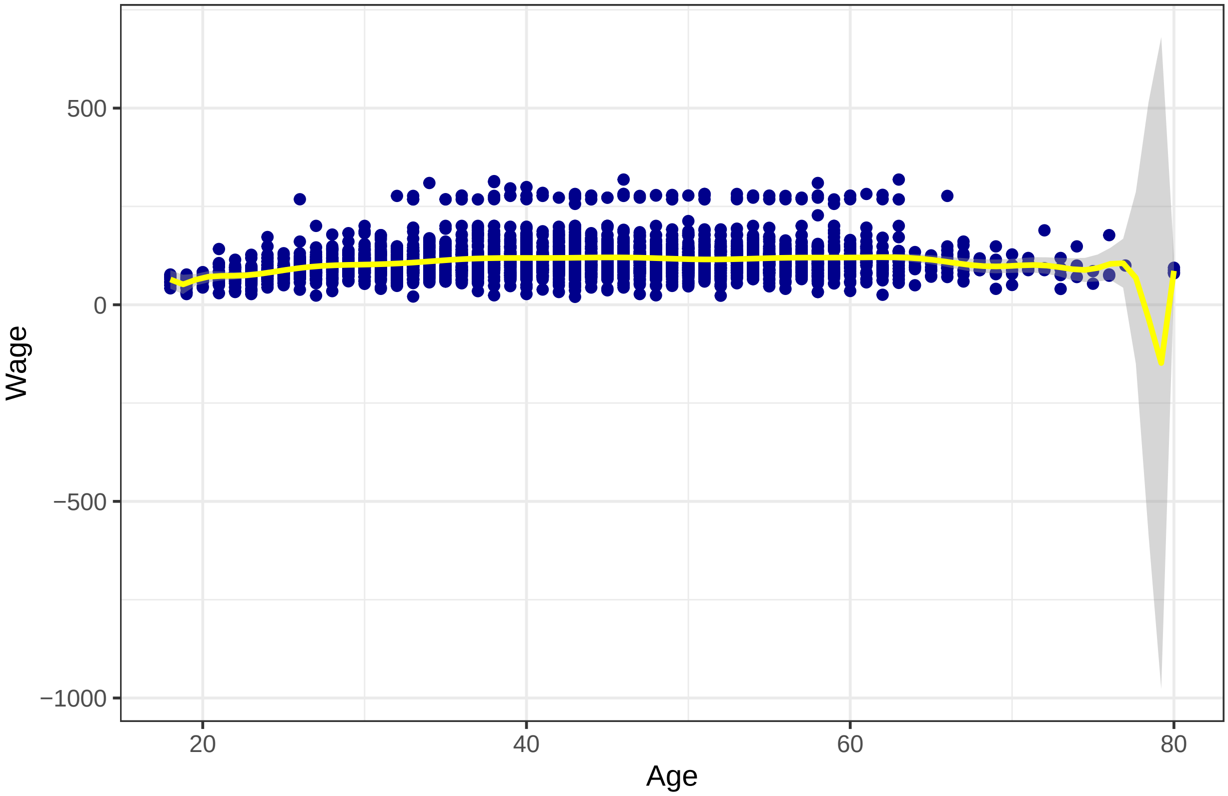


fitting polynomial regression.





Polynomial degree 5



Polynomial degree 20

```
##    Degree of the age polynomial R-Squared
## 1                             2 0.2896871
## 2                             3 0.2908565
## 3                             4 0.2908565
## 4                             5 0.2914362
## 5                             6 0.2918935
## 6                             7 0.2928255
## 7                             8 0.2928256
## 8                             9 0.2935562
## 9                            10 0.2937707
## 10                           11 0.2937954
## 11                           12 0.2937982
## 12                           13 0.2938966
## 13                           14 0.2940063
## 14                           15 0.2941473
## 15                           16 0.2942057
## 16                           17 0.2947922
## 17                           18 0.2947927
## 18                           19 0.2948218
## 19                           20 0.2948309
```

Even the higher degree give the higher R^2, the `overfitting` problem may be occured. Hence, polynomial regression with degree 3 would be appropriate.

```r
model_poly <- lm(wage ~ poly(age, 3) + education + year, data = data)
summary(model_poly)
```

```
##
## Call:
## lm(formula = wage ~ poly(age, 3) + education + year, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -118.565  -19.789   -3.339   14.399  213.276
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -2247.6445   637.2171  -3.527 0.000426 ***
## poly(age, 3)1                 358.1166    35.4147  10.112  < 2e-16 ***
## poly(age, 3)2                -383.1188    35.3679 -10.832  < 2e-16 ***
## poly(age, 3)3                  78.2802    35.2489   2.221 0.026440 *
## education2. HS Grad            10.8127     2.4290   4.452 8.84e-06 ***
## education3. Some College       23.2840     2.5564   9.108  < 2e-16 ***
## education4. College Grad       37.8823     2.5414  14.906  < 2e-16 ***
## education5. Advanced Degree    62.4402     2.7584  22.636  < 2e-16 ***
## year                            1.1633     0.3177   3.662 0.000255 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.19 on 2991 degrees of freedom
## Multiple R-squared:  0.2909, Adjusted R-squared:  0.289
## F-statistic: 153.3 on 8 and 2991 DF,  p-value: < 2.2e-16
```

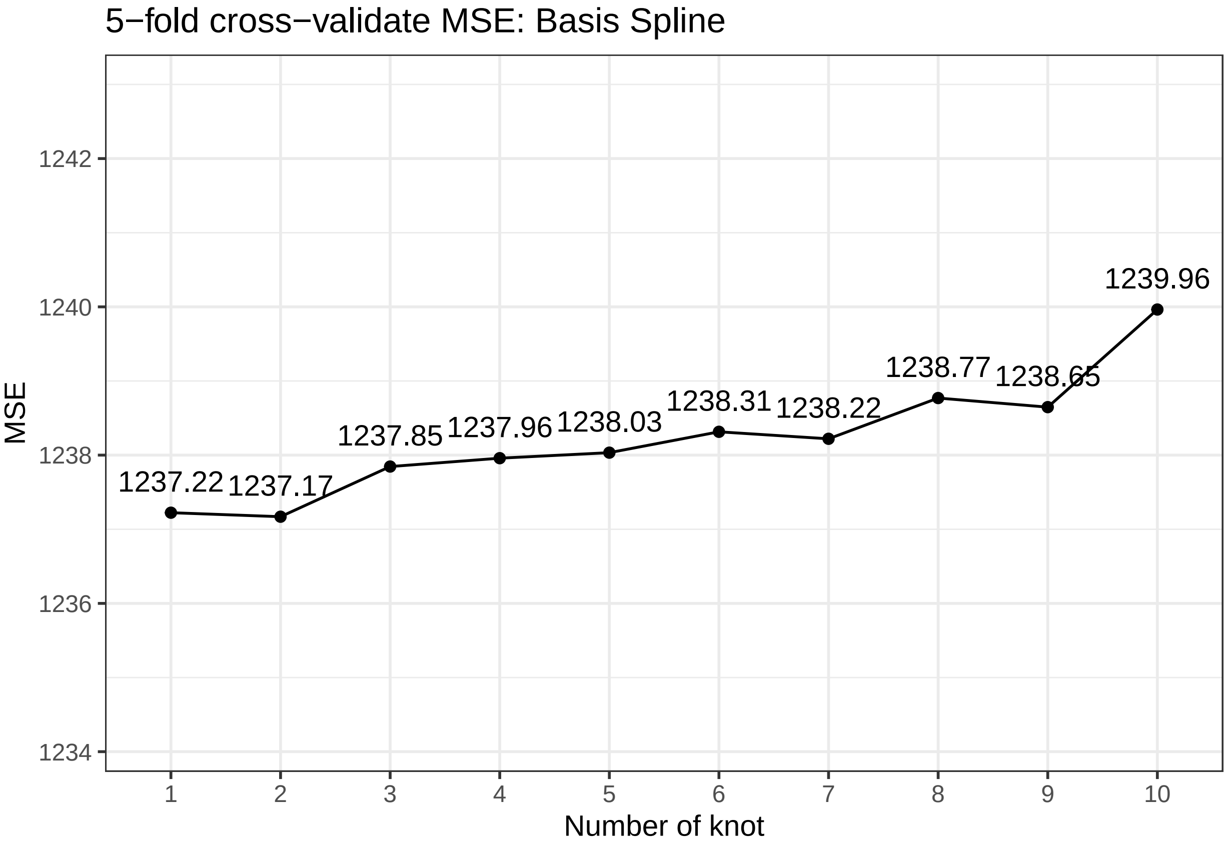
The R^2 is 0.2909, which improved from all previous models.

Basis Spline and Natural Spline

For both `Basis Spline` and `Natural Spline`, the number of knots or the degree of freedom need to be specified. One of the method used for specified is performing `K-fold Cross Validation`. In this case, K is equal to 5. For both types of spline, the highest degree of polynomial for age is 3.

- Basis Spline: df = 4 + knots
- Natural Spline: df = 2 + knots

Consider the MSE for basis spline.



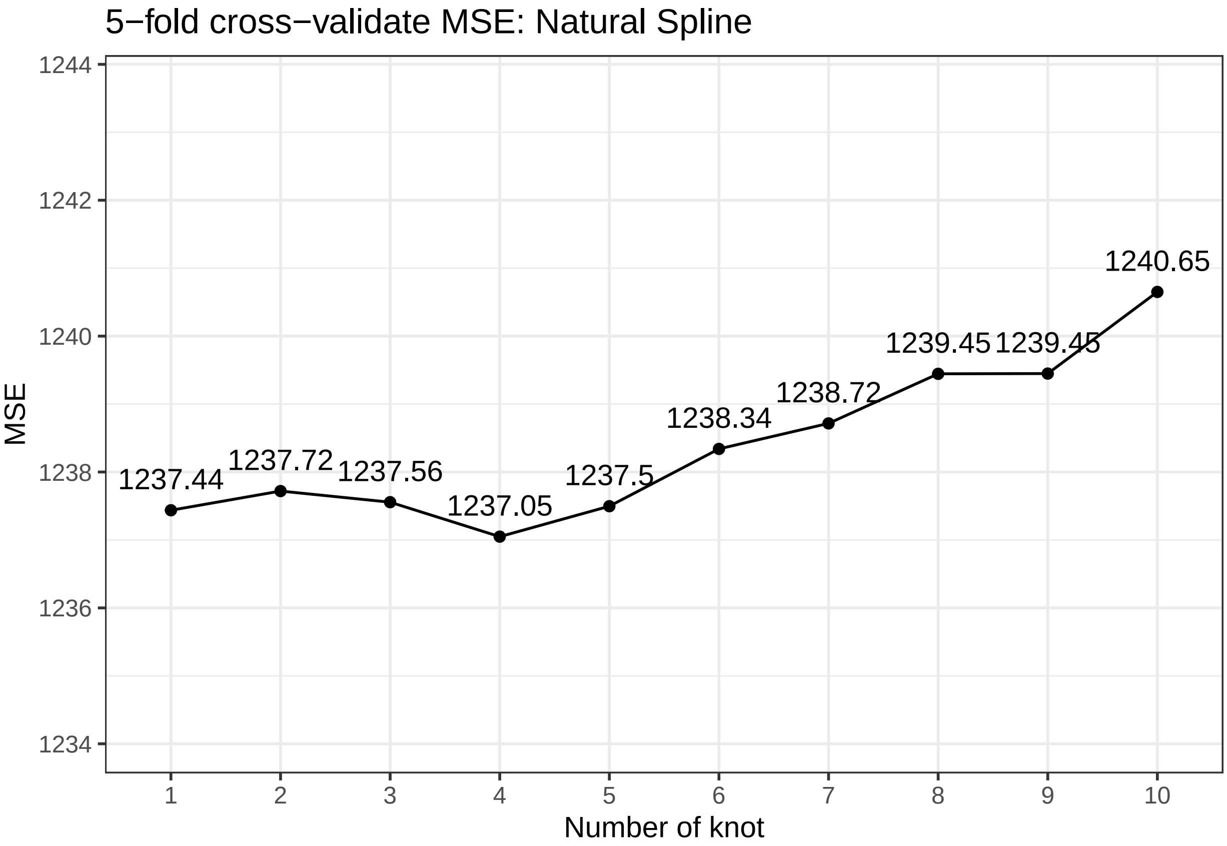
The MSE is lowest when the number of `knot` is equal to 2. Fit the regression with basis spline.

```
model_basis <- lm(wage ~ bs(age, df = 6) + education + year, data = data)
summary(model_basis)
```

```
##
## Call:
## lm(formula = wage ~ bs(age, df = 6) + education + year, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -120.371  -19.640   -3.273   14.086  213.170
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -2344.664    637.830  -3.676 0.000241 ***
## bs(age, df = 6)1                11.675     10.997   1.062 0.288473
## bs(age, df = 6)2                31.678      6.333   5.002 6.01e-07 ***
## bs(age, df = 6)3                46.964      7.371   6.372 2.16e-10 ***
## bs(age, df = 6)4                34.013      7.742   4.393 1.16e-05 ***
## bs(age, df = 6)5                 48.731      12.143   4.013 6.14e-05 ***
## bs(age, df = 6)6                  6.633      14.292   0.464 0.642610
## education2. HS Grad              11.075       2.430   4.557 5.41e-06 ***
## education3. Some College         23.638       2.562   9.227  < 2e-16 ***
## education4. College Grad         38.242       2.548  15.008  < 2e-16 ***
## education5. Advanced Degree      62.597       2.761  22.669  < 2e-16 ***
## year                              1.194       0.318   3.753 0.000178 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.17 on 2988 degrees of freedom
## Multiple R-squared:  0.2923, Adjusted R-squared:  0.2897
## F-statistic: 112.2 on 11 and 2988 DF,  p-value: < 2.2e-16
```

The R^2 is 0.2923.

Then consider the Natural Spline.



The MSE is lowest when the number of `knot` is equal to 4. Fit the regression with natural spline.

```
model_natural <- lm(wage ~ ns(age, df = 6) + education + year, data = data)
summary(model_natural)
```

```
##
## Call:
## lm(formula = wage ~ ns(age, df = 6) + education + year, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -121.403  -19.727   -3.143   14.174  214.340
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                 -2394.4450   638.1274  -3.752 0.000179 ***
## ns(age, df = 6)1               38.7338     4.6496   8.331  < 2e-16 ***
## ns(age, df = 6)2               46.4652     5.8970   7.879 4.57e-15 ***
## ns(age, df = 6)3               38.1178     5.1218   7.442 1.29e-13 ***
## ns(age, df = 6)4               37.0673     4.8062   7.712 1.67e-14 ***
## ns(age, df = 6)5               48.9899    11.6639   4.200 2.75e-05 ***
## ns(age, df = 6)6                4.3620     8.9214   0.489 0.624922    
## education2. HS Grad            11.1264     2.4295   4.580 4.85e-06 ***
## education3. Some College       23.6491     2.5595   9.240  < 2e-16 ***
## education4. College Grad       38.3108     2.5454  15.051  < 2e-16 ***
## education5. Advanced Degree    62.5971     2.7605  22.676  < 2e-16 ***
## year                            1.2186     0.3182   3.830 0.000131 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.16 on 2988 degrees of freedom
## Multiple R-squared:  0.2927, Adjusted R-squared:  0.2901
## F-statistic: 112.4 on 11 and 2988 DF,  p-value: < 2.2e-16
```

The R^2 is 0.2927.

Summary

Discussion

Reference