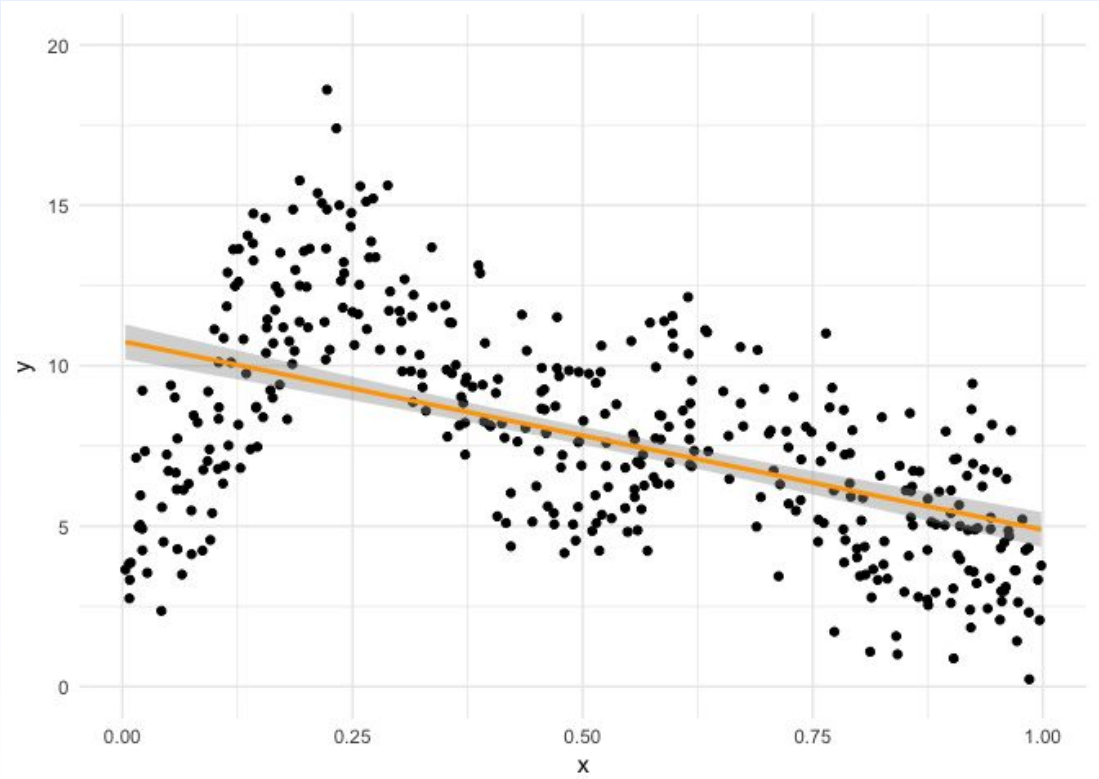
Generalized Additive Model (GAM)

By Cole Conte, Jiayou Chao, Peter Gonatas, Steven Geiser

Background

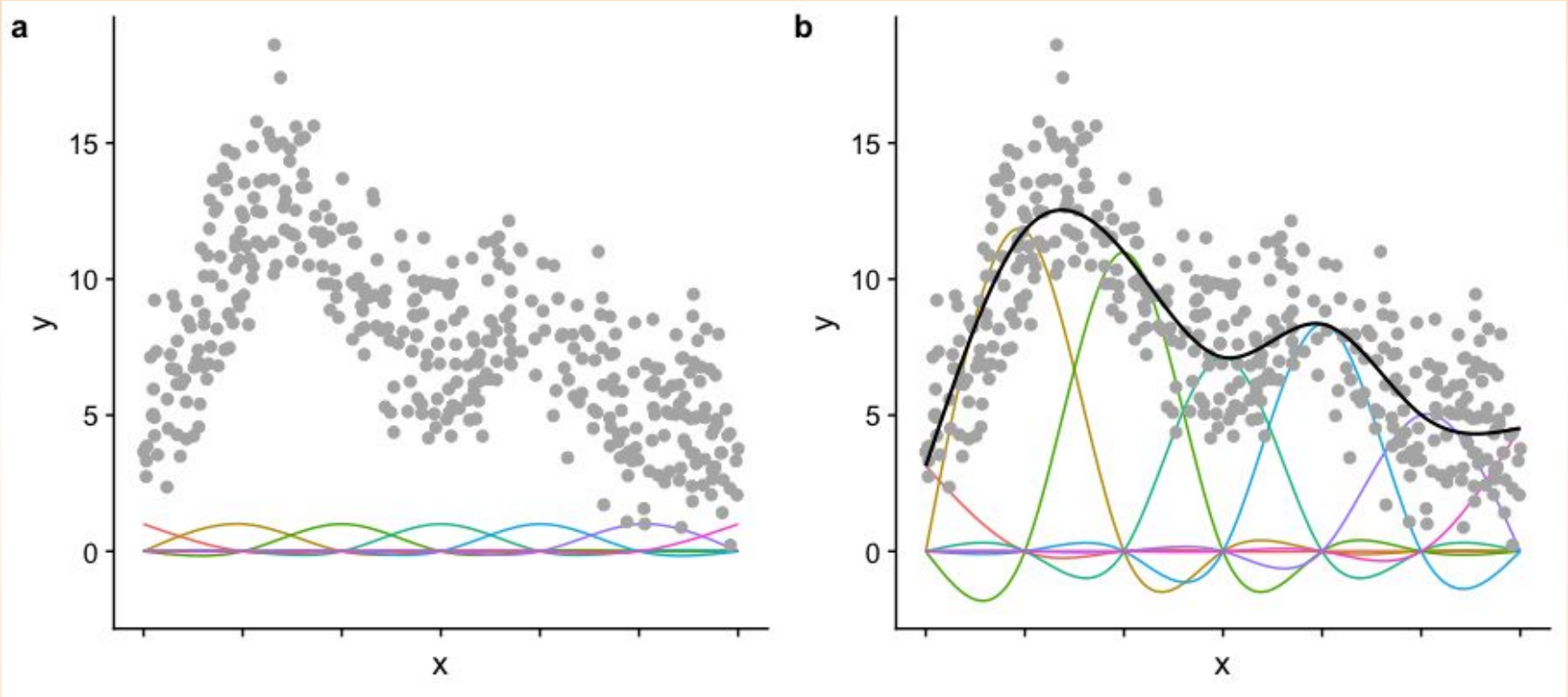
- The main feature that distinguishes the GAM from other models is that each predictor variable is assigned its own smoothing function and these functions are added together to create a model.
- Each function is also nonparametric, meaning that they are shaped by the data alone rather than parameters of previously defined distributions. Despite this, these functions can be both linear and nonlinear. This allows GAM to be a lot more flexible than other models.
- One of the advantages of GAM is that because of how you set up the smoothing functions, you can avoid over/underfitting the data.
- GAM can also support the use of link functions like logit.

Linear Model



Source: https://noamross.github.io/gams-in-r-course/chapter1

GAM

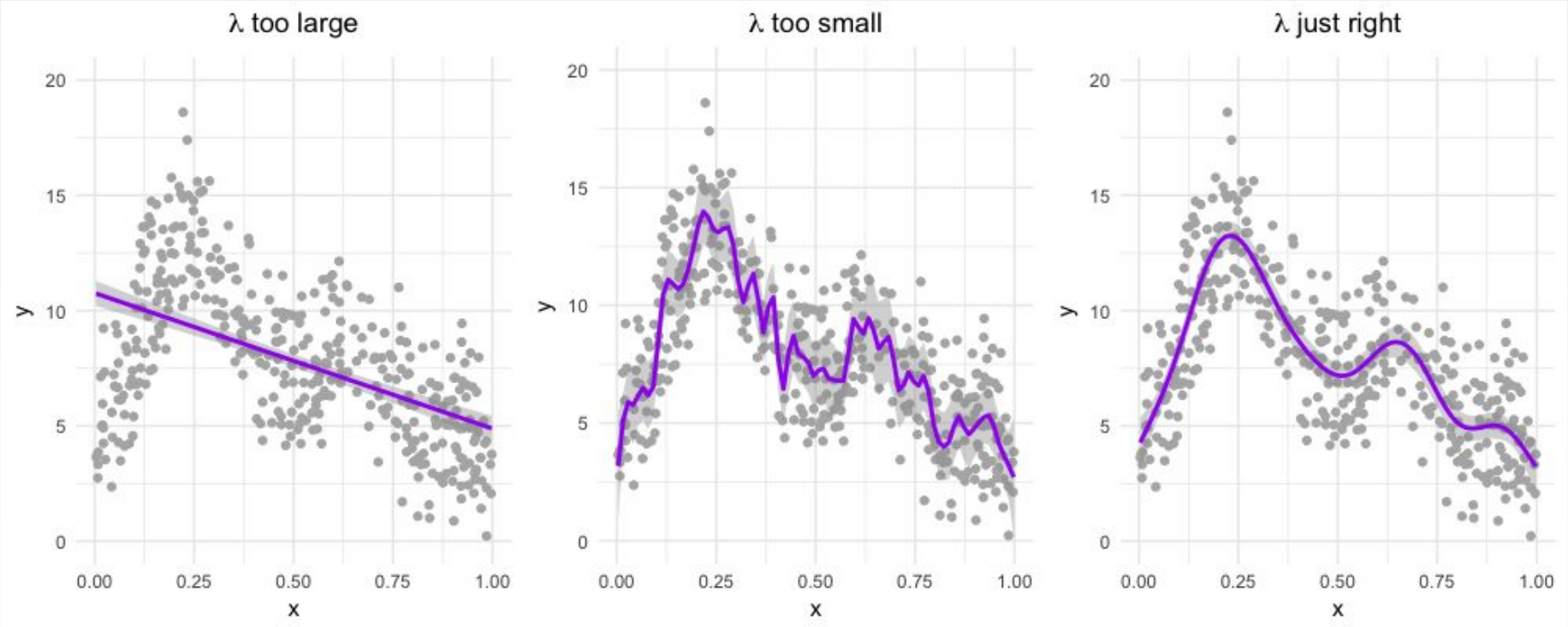


Source: https://noamross.github.io/gams-in-r-course/chapter1

Takeaways

- As highlighted by the previous graphs, there is a significant benefit to using GAM over regular linear models.
- When your data follows an irregular/ nonlinear pattern, a linear model may represent an overall trend, but cannot accurately predict a trend over a smaller section.
- Using GAM a combination smoothing function can be used to more accurately fit a line to a complex data set.

Smoothing Parameter

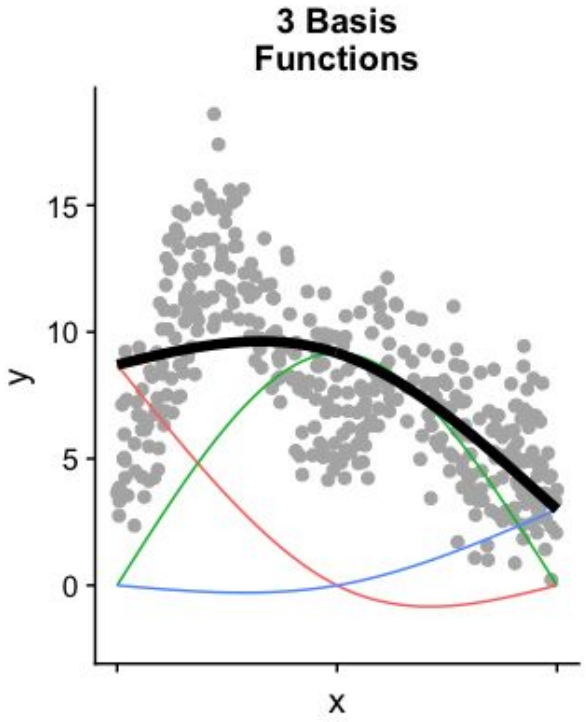


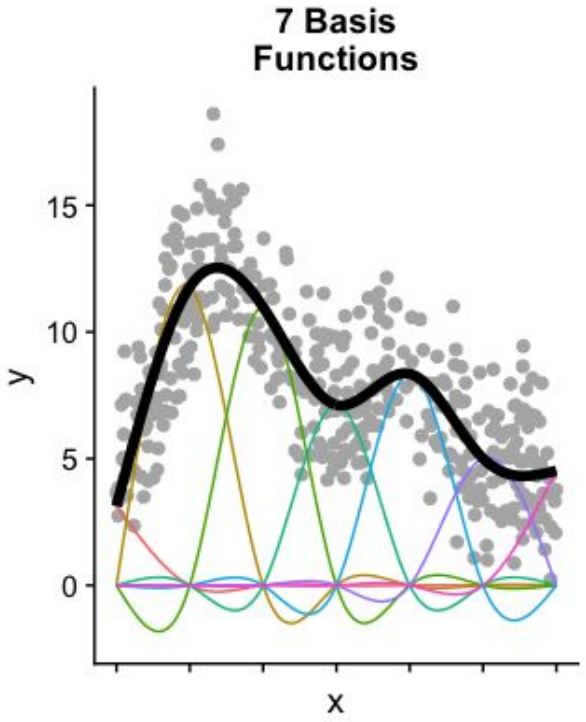
Source: https://noamross.github.io/gams-in-r-course/chapter1

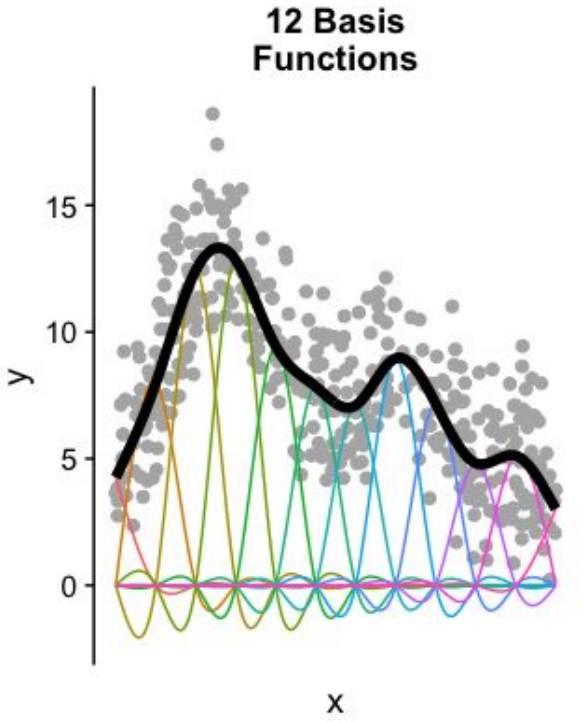
Smoothing Parameter

- When trying to fit the data pattern, the smoothing parameter λ is important for making sure our function fits.
- As shown in the previous graphs, if the smoothing parameter is too large, it will underfit the graph and behave similarly to a linear function. If it is too small, the smoothing function will overfit the data and becomes over sensitive to small changes in the data. Therefore, it is important to pick a smoothing parameter that bests fits the data.
- When using R, we will use method = "REML" to decide the smoothing parameter.

Basis Functions







Source: https://noamross.github.io/gams-in-r-course/chapter1

Basis Functions

- The basis functions are the individual functions combined to make a smoothing function.
- Like the smoothing parameter, the smoothing function can be over/under-fitted depending on the number of basis functions, so it is important to balance the number you use.

Model Fitting

Much like Ordinary Least Squares with traditional models, the central component of fitting a GAM is an optimization problem. This penalized least squares problem aims to find the function that minimizes the following:

$$\sum_{i=1}^{n}\{Y_i - \hat{f}(x_i)\}^2 + \lambda \int \hat{f}''(x)^2\, dx.$$

Sum of Squared Errors

Smoothing Parameter

Roughness (Measured by the integral of the square of the second derivative)

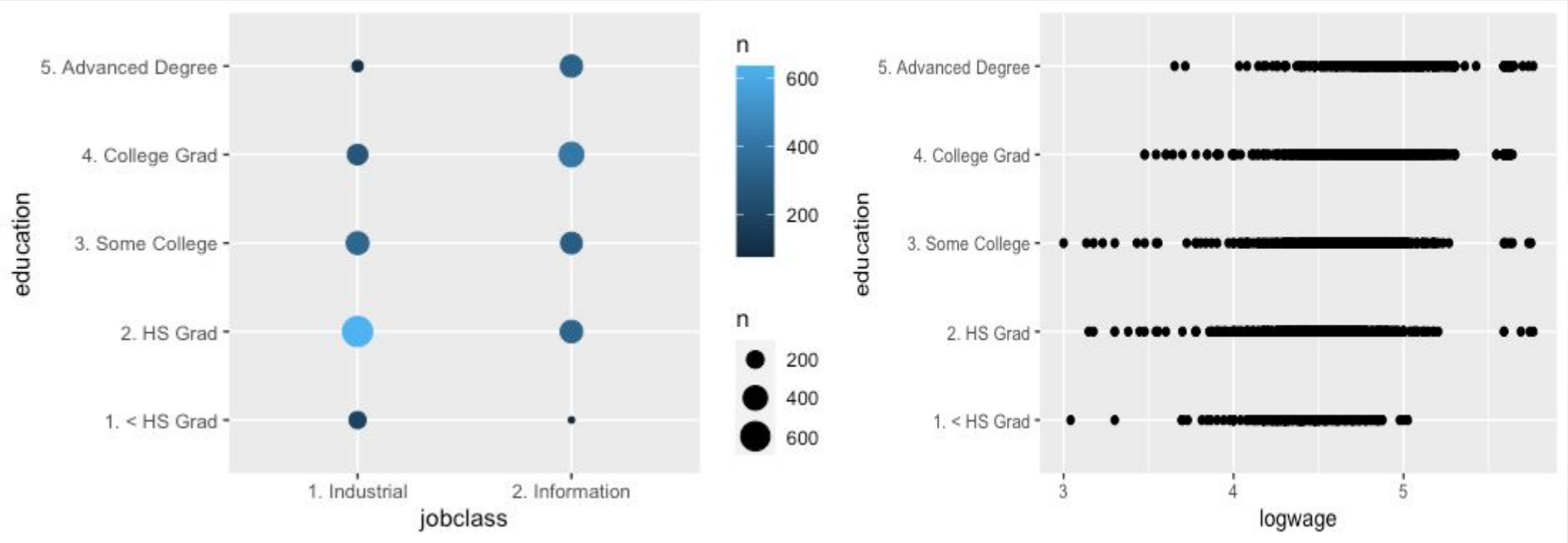
There are several methods we can use to approximate this optimal function from a given number of basis functions. For our data, we will primarily focus on fitting via thin-plate regression splines. These splines are a low-rank approximation to a “full” spline formed from n-1 basis functions, where n is the total number of observations.

Our Dataset

The dataset we used to fit our models contains data from 3,000 male workers in the Mid-Atlantic region. The variables from the data include:

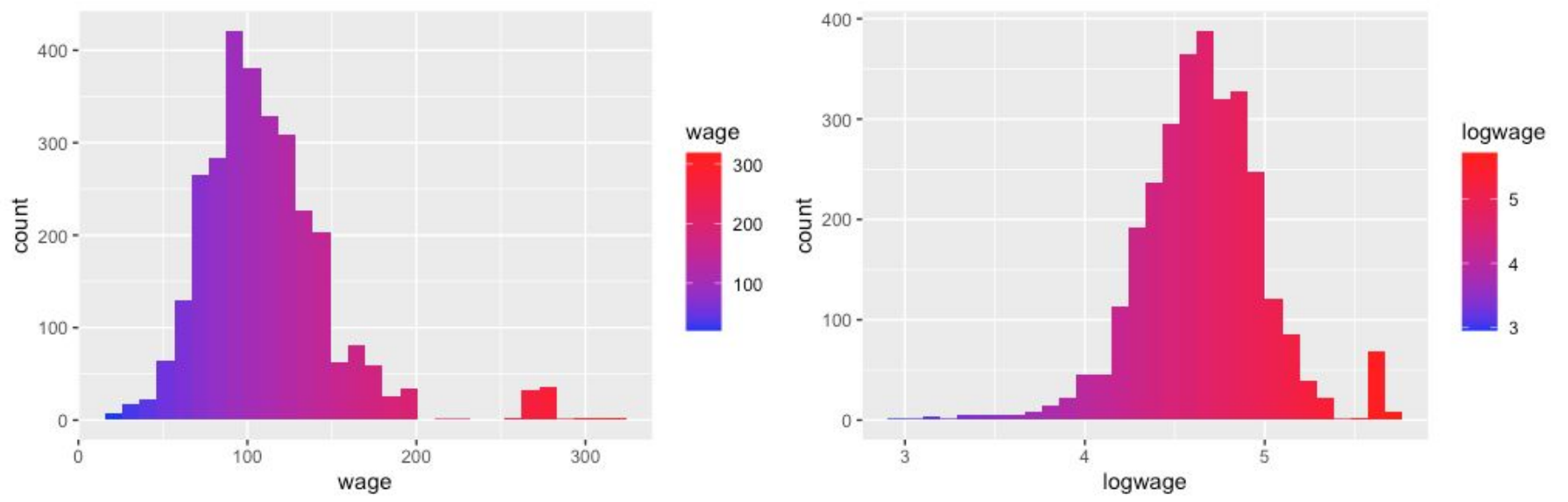
- Year: Year of recording the data
- Age: Age of the worker in years
- Maritl: Categorical- Never Married, Married, Widowed, Divorced, Separated (5 Categories)
- Race: Categorical- White, Black, Asian, Other (4 Categories)
- Education: Categorical (Ordinal)- <HS Grad, HS Grad, Some College, College Grad, Advanced Degree (5 Categories)
- Region: Categorical- The worker's region, Mid-Atlantic for all observations (1 Category)
- Jobclass: Categorical- Type of job, industrial or information (2 Categories)
- Health: Categorical- The worker's health, split into ≤ Good and ≥ Very Good (2 Categories)
- Health_ins: Categorical - Indicates whether the worker has health insurance (2 Categories)
- logwage: The natural logarithm of the worker's wage
- wage: The worker's wage, in thousands of U.S. dollars ($)

Observations from EDA

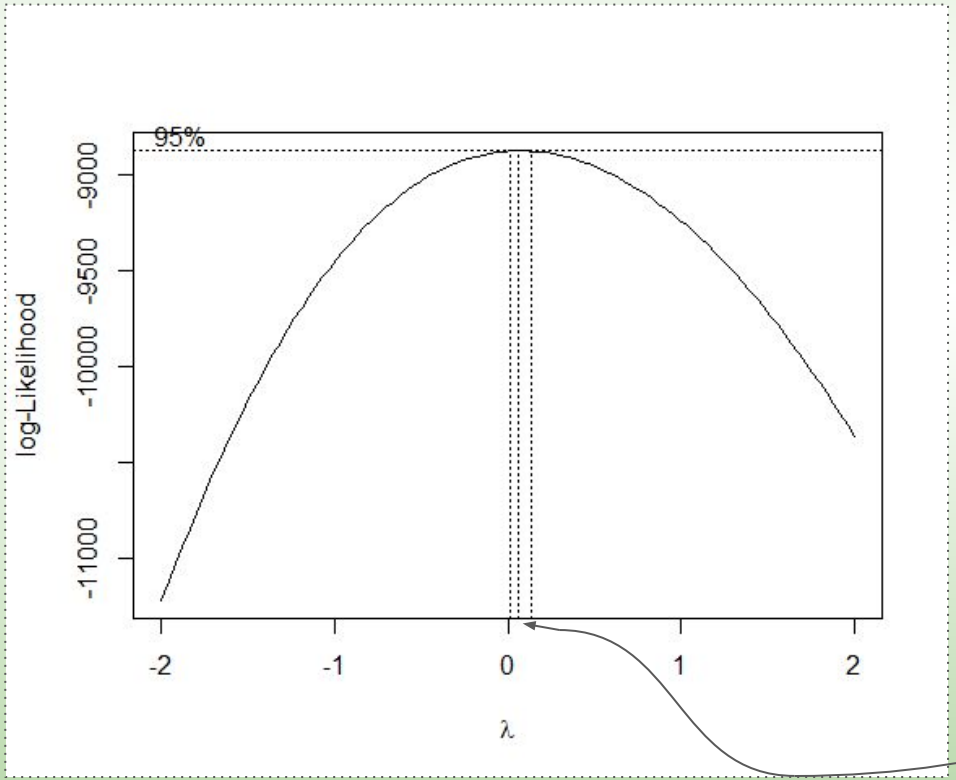


Transformation of Wage

In order to mitigate the influence of outliers, we chose to use the log-transformation of wage in our model instead. The original distribution of the wage variable is heavily right-skewed with several outliers, so this transformation would prove beneficial to our analysis.



Box-Cox Wage Transformation



With a high log-likelihood for λ=0, a logarithmic transformation is a reasonable choice for a roughly symmetric distribution of wage values.

Fitting a Simple GAM

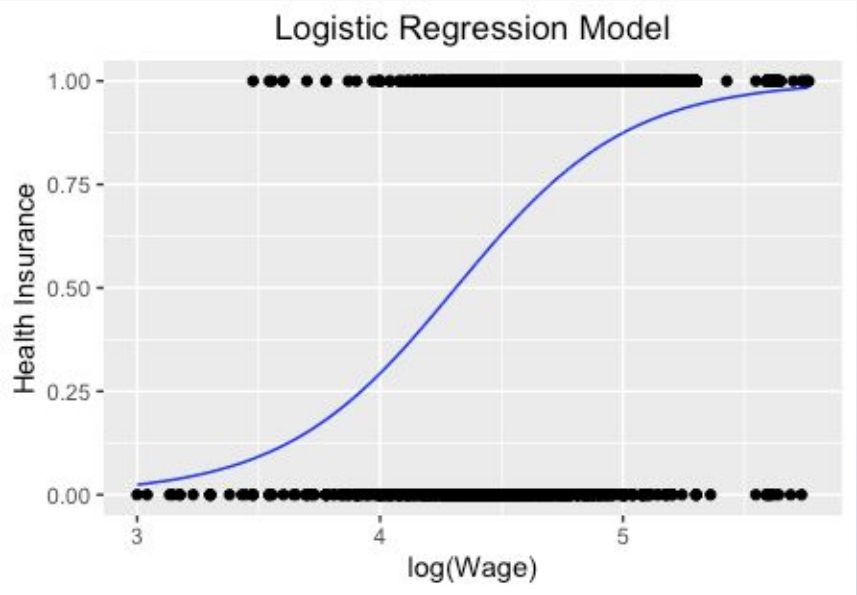
We would like to find a proper model to predict whether a worker has health insurance based on wage. The style of this GAM will be similar to that of single-predictor logistic regression, using the logit link function and specifying health insurance as having an underlying binomial distribution:

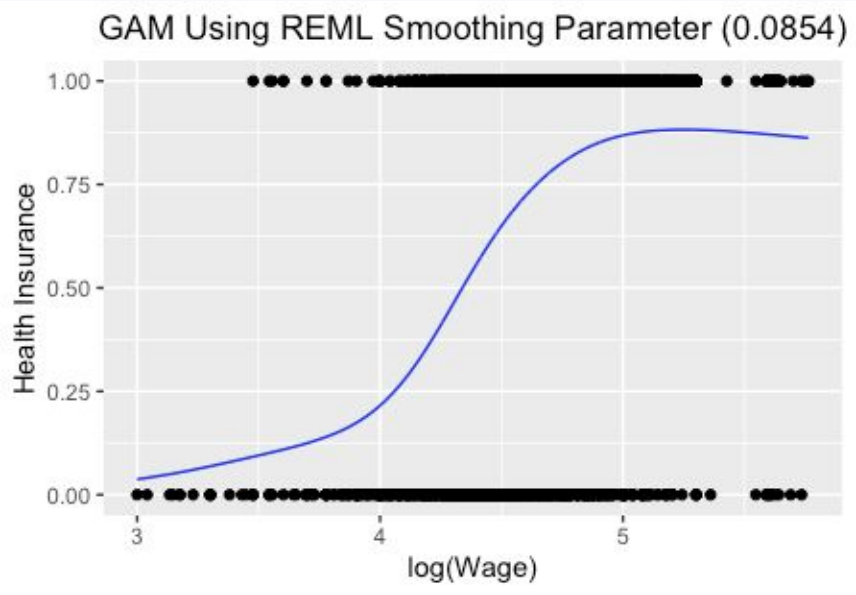
$$\text{logit}[E(Y)] = s_0 + s_1(X_1)$$

Where Y is a binary variable equal to 1 if a worker has health insurance and 0 otherwise, and X_1 denotes wage. s_0 is the constant term and s_1 represents a smooth function composed of several basis functions used to fit the relationship between the dependent and independent variable. Each smoothing can actually be expressed as a linear combination of these functions.

GLM vs. GAM: Plot Comparison

We can fit our model via logistic regression or use a generalized additive model for a nonparametric approach. In comparing the regression functions for these 2 cases, we notice key similarities and distinctions between them. While both models maintain the same general shape, we can see that the GAM, which uses 9 basis functions, depicts a nonlinear relationship in which the dependent variable is not strictly increasing.





R Output

Logistic Regression

```
Call:
glm(formula = health_ins ~ logwage, family = binomial)

Coefficients:
             Estimate  Std. Error z value  Pr(>|z|)
(Intercept) -12.1661     0.7013  -17.35   <2e-16 ***
logwage       2.8217     0.1532   18.41   <2e-16 ***

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3693.5  on 2999  degrees of freedom
Residual deviance: 3240.2  on 2998  degrees of freedom

AIC: 3244.246
BIC: 3256.259
```

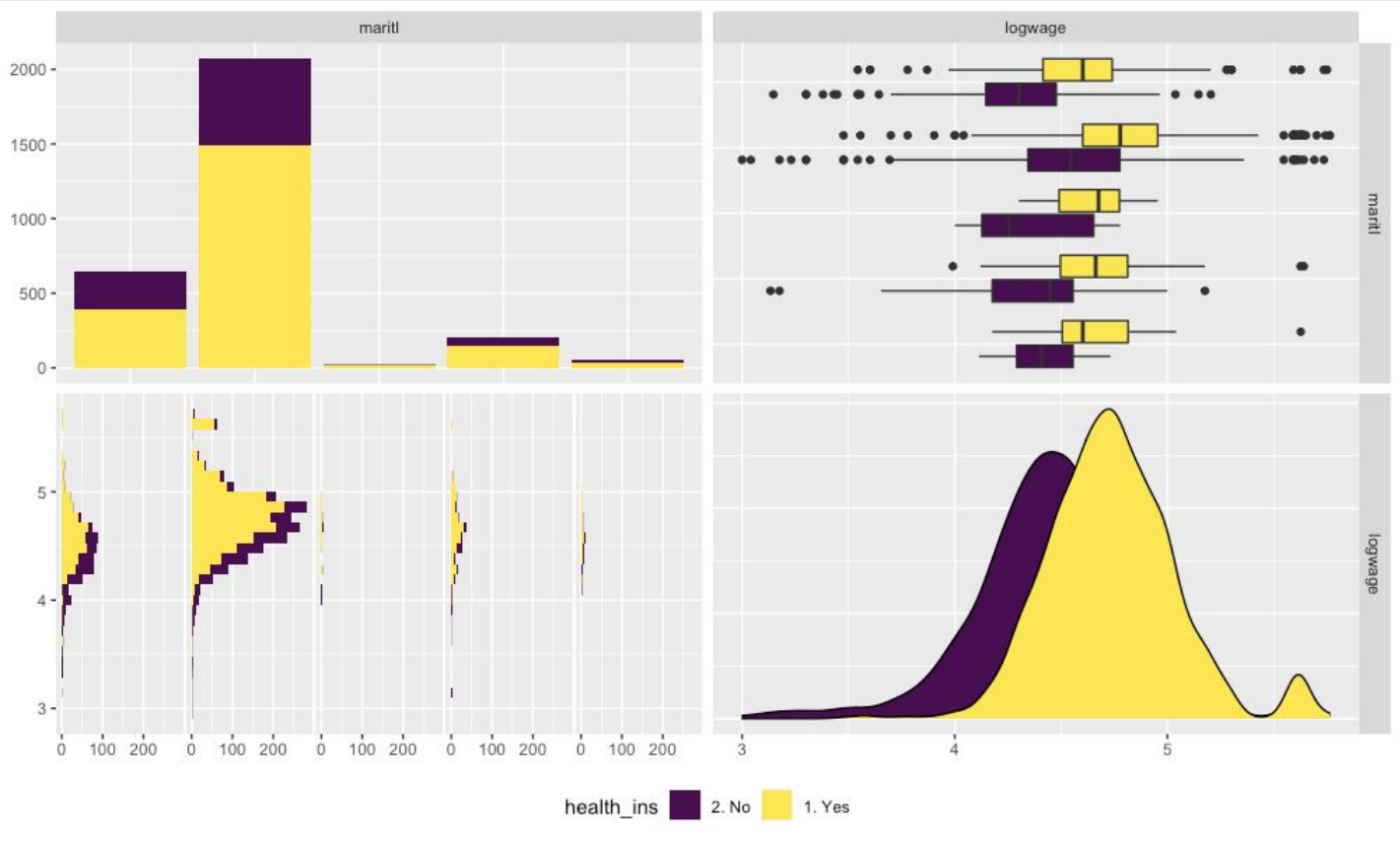
GAM

```
Family: binomial
Link function: logit
Parametric coefficients:
             Estimate  Std. Error z value  Pr(>|z|)
(Intercept)  0.92462    0.04526   20.43   <2e-16 ***

Approximate significance of smooth terms:
               edf    Ref.df Chi.sq  p-value
s(logwage) 4.527  5.575  396.2  <2e-16 ***

R-sq.(adj) =  0.167   Deviance explained = 13.4%
-REML = 1608.8  Scale est. = 1         n = 3000
Residual deviance: 3197.48
AIC: 3209.386
BIC: 3245.141
```

The edf term in the GAM output is the estimated degrees of freedom. It can be interpreted as a measure of complexity for a variable's fitted spline: An edf of 1 corresponds to a linear fit, and higher values indicate complexity akin to higher-order polynomials. While both models convey significant results for the coefficient(s) of logwage, the latter model may be preferred based on diagnostics such as deviance and AIC/BIC.

Category-Level Smoothing

When dealing with categorical variables, we have the option to apply a smooth term to a continuous variable that depends on a categorical variable. For instance, using our simple GAM from before, we can apply a smoothing function to log(wage) by marital status.



R Output

GAM by Category

```
Family: binomial
Link function: logit
Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.94475    0.04737   19.94   <2e-16 ***

Approximate significance of smooth terms:
                                    edf    Ref.df  Chi.sq    p-value
s(logwage):maritl1. Never Married 3.599  4.539 149.562  < 2e-16 ***
s(logwage):maritl2. Married       3.761  4.711 211.861  < 2e-16 ***
s(logwage):maritl3. Widowed       2.307  2.887   3.086 0.306069
s(logwage):maritl4. Divorced      2.333  3.004  22.251 5.81e-05 ***
s(logwage):maritl5. Separated     1.001  1.002  12.486 0.000413 ***

R-sq.(adj) =  0.168   Deviance explained = 13.8%
-REML = 1608.8  Scale est. = 1         n = 3000
Residual deviance: 3184.62
AIC: 3218.905
BIC: 3321.873
```

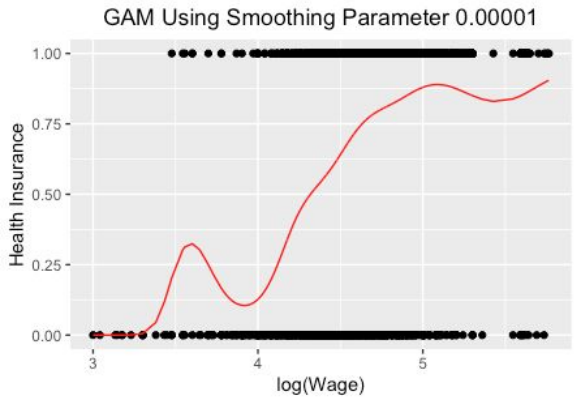
GAM

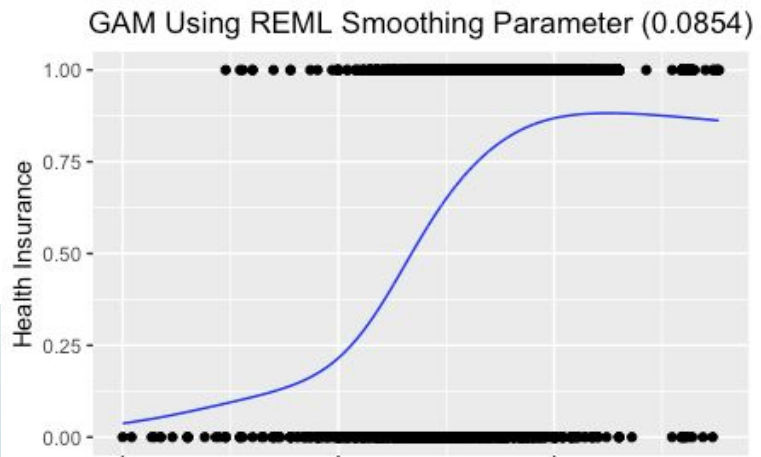
```
Family: binomial
Link function: logit
Parametric coefficients:
            Estimate  Std. Error z value  Pr(>|z|)
(Intercept)  0.92462    0.04526   20.43   <2e-16 ***

Approximate significance of smooth terms:
               edf    Ref.df Chi.sq  p-value
s(logwage) 4.527  5.575  396.2  <2e-16 ***

R-sq.(adj) =  0.167   Deviance explained = 13.4%
-REML = 1608.8  Scale est. = 1         n = 3000
Residual deviance: 3197.48
AIC: 3209.386
BIC: 3245.141
```

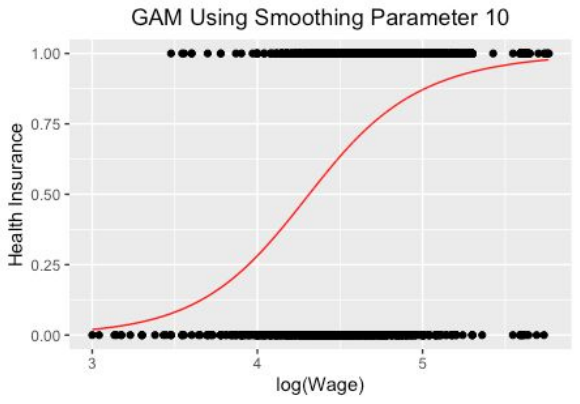
Now we have 5 different smoothed predictors: one for each level of marital status.

Changing the Smoothing Parameter (λ)

We chose to fit our initial GAM using Restricted Maximum Likelihood (REML), which estimates the “optimal” smoothing parameter for our model. As $\lambda \to \infty$, our model will converge to the linear case. In contrast, as $\lambda \to 0$, our model will become less smooth and its behavior will be inconsistent (tending to overfit the data). Models using different values of the smoothing parameter are shown below:

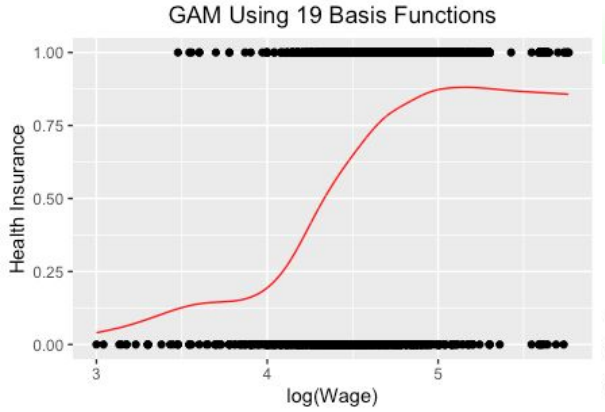


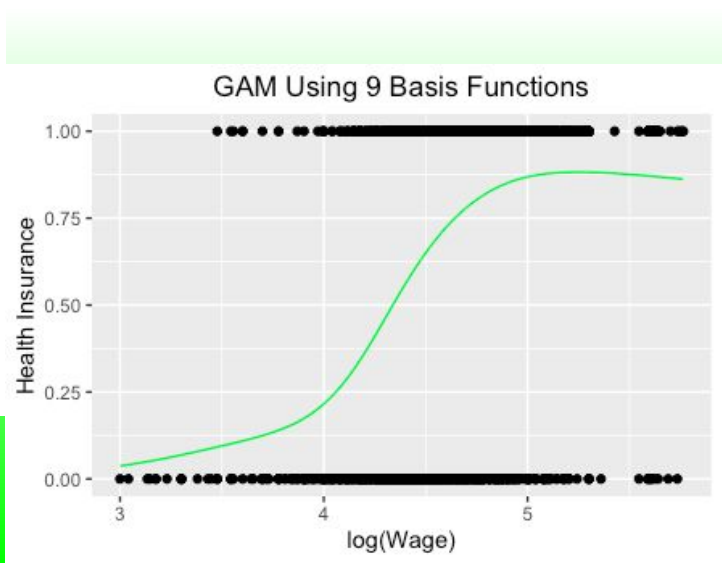


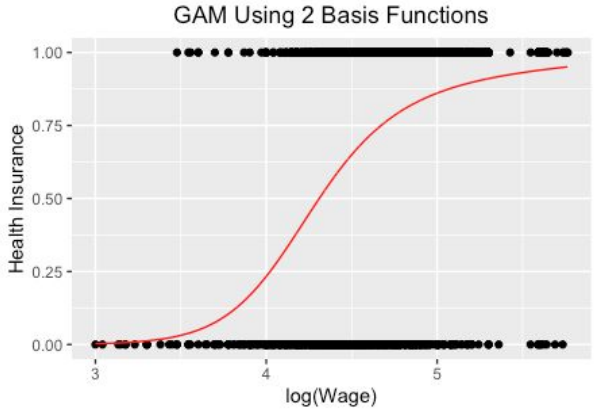


Changing the Number of Basis Functions

As mentioned before, the number of basis functions is the other attribute used to find the appropriate balance between accuracy and overfitting. By default, our single-predictor model used 9 basis functions. As the number of functions decreases, our model will be more simplistic with fewer parameters. On the other hand, a regression with many basis functions will be more sensitive to trends in the data. Below are models using different numbers of basis functions, using the same smoothing parameter:

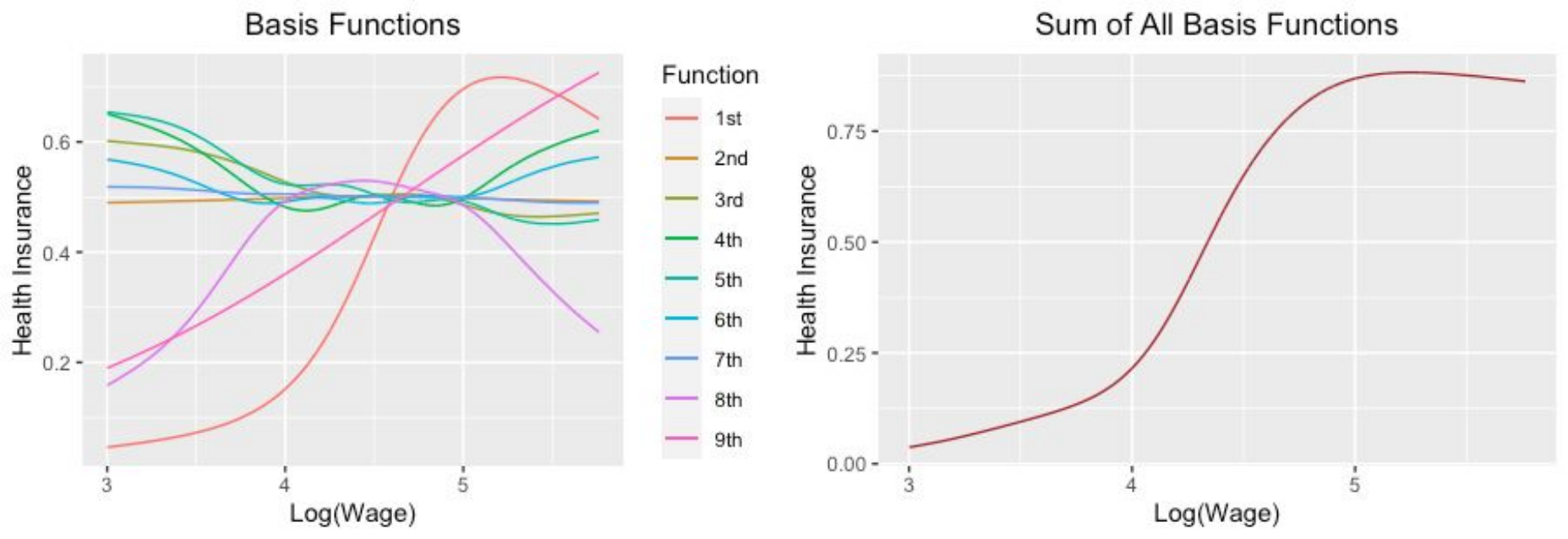






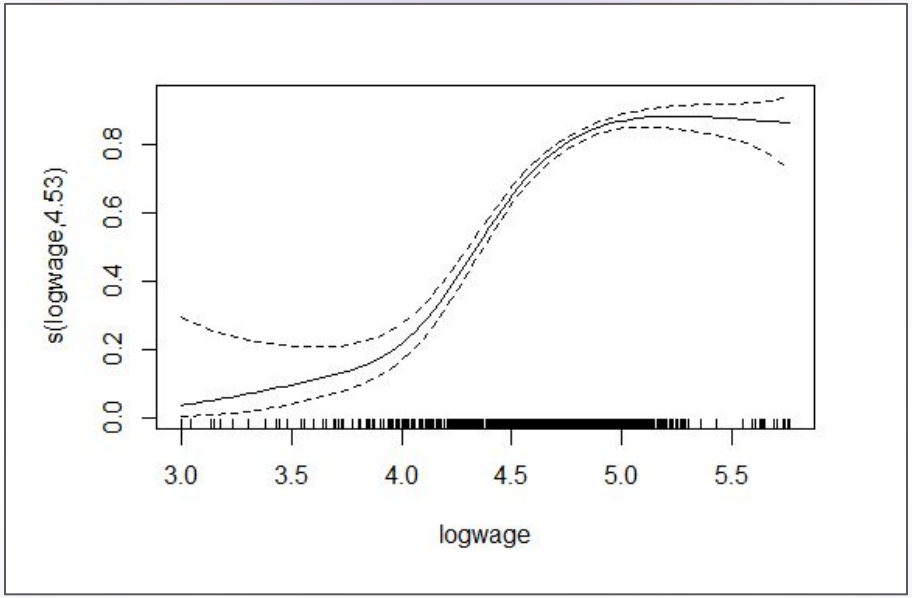
Visualizing our Basis Functions

Here we can visualize each individual basis function in our 9 basis function GAM, as well as the basis functions summed together (which is just our original fit). After taking the inverse logit-transform of this sum, we will end up with the same estimated curve for our model.



Assessing the Adequacy of Our GAM

Subjectively, a GAM is deemed “adequate” if it uses a proper number of basis functions to capture the major trends in the data. We must remind ourselves that we want our model to have enough parameters to identify these trends, but models built from more basis functions are more loosely bounded in terms of maximum “wiggliness”. Does our 9-function model suffice?



Model Testing

To answer our question, we examine 2 aspects of our model: convergence and the distribution of residuals. If the algorithm for maximizing likelihood function for the model coefficients does not converge, the GAM is likely not adequate. This lack of convergence is often attributed to an overabundance of parameters. Next, we test if the distribution of residuals is random. If the p-value for this test is significant, we conclude that the residuals are not random. In this case, too few basis functions is the most likely cause.

Using R:

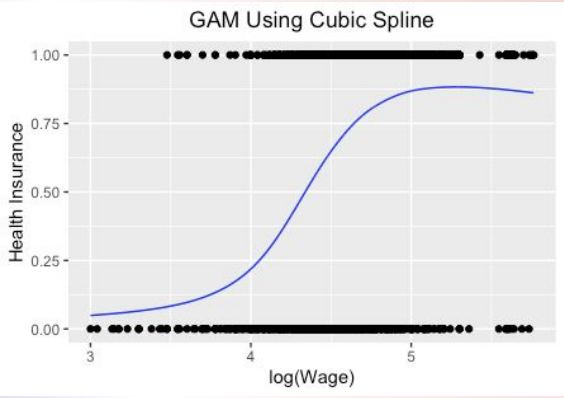
```
Method: REML   Optimizer: outer newton
full convergence after 3 iterations.
Gradient range [1.182873e-06,1.182873e-06]
(score 1608.838 & scale 1).
Hessian positive definite, eigenvalue range [1.348296,1.348296].
Model rank =  10 / 10

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                 k'   edf  k-index p-value
s(logwage) 9.00 4.53    0.97    0.18
```

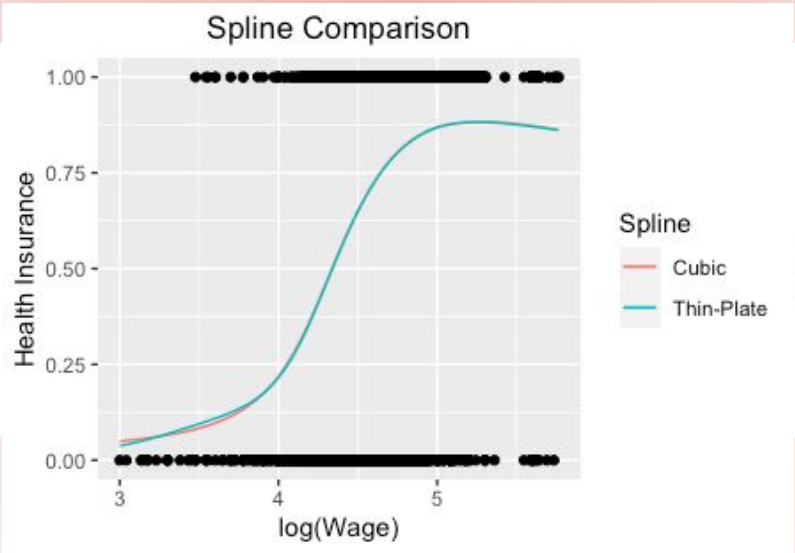
With convergence achieved using Newton's method and a p-value displaying no significance in the test for randomized residuals, it appears that this fitted model performs adequately for the data.

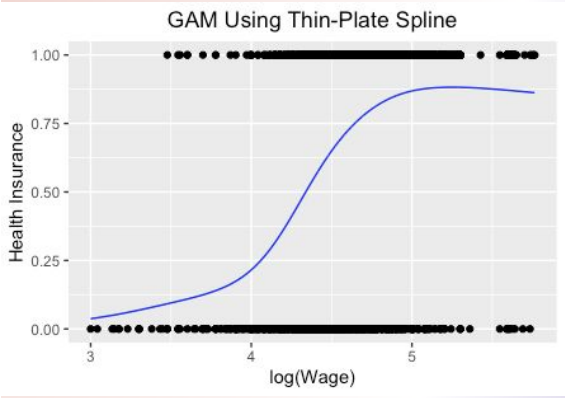
Although the k-index is less than 1, the estimated degrees of freedom is not close to the number of basis functions (k).

Thin-Plate and Cubic Splines

In our earlier example, we used thin-plate splines (TPS) to construct our GAM. We can also re-create our model using the traditional cubic spline. Both methods produce a piecewise curve joined at knots, but only thin-plate splines select the locations in the process. For comparison, below are the original curve using thin-plate splines and the curve made from a cubic spline using evenly-spaced knots. In this case, the results are quite comparable:

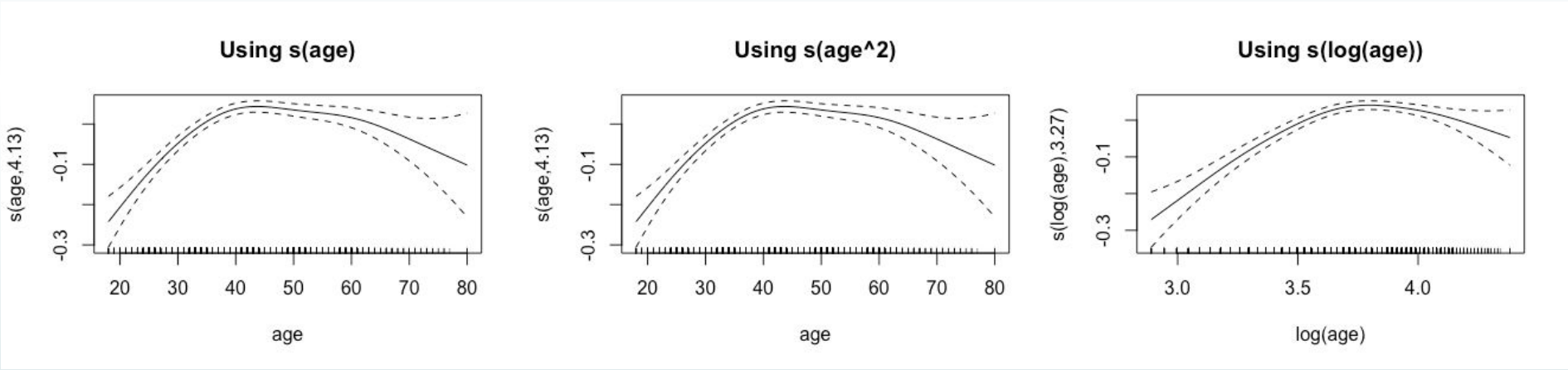






Multivariate GAM

Here we use log(Wage) as the dependent variable and try to build a multivariate GAM. In other words, the independent variables are year, age, maritl, education, jobclass, health and health_ins, among which age is a continuous variable that we smooth while others are categorical variables. The goal is to explain as much deviance as possible.



Multivariate GAM

The difference of the three models is that we use ‘age’ as the predictor in the first model, ‘age^2’ in the second model and ‘log(age)’ in the third model. However, as we can see, the adjusted R-squared (0.387) the deviance explained (39.2%) remain the same.

Age

```
Approximate significance of smooth terms:
         edf Ref.df    F p-value
s(age) 4.131  5.128 19.8  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05

R-sq.(adj) =  0.387   Deviance explained = 39.2%
```

Age^2

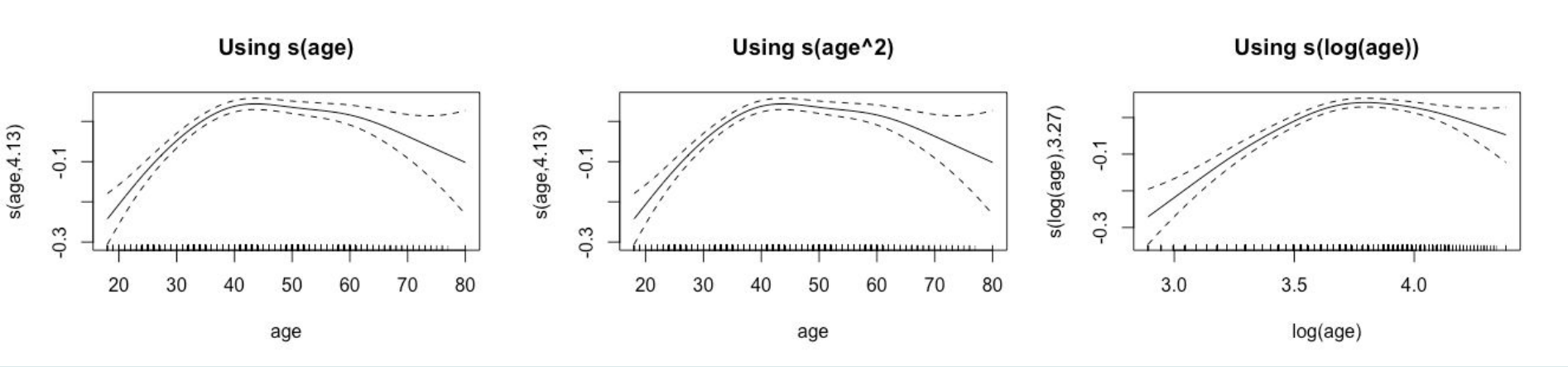
```
Approximate significance of smooth terms:
         edf Ref.df    F p-value
s(age) 4.131  5.128 19.8  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05

R-sq.(adj) =  0.387   Deviance explained = 39.2%
```

log(Age)

```
Approximate significance of smooth terms:
             edf Ref.df     F p-value
s(log(age)) 3.268  4.121 24.21  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘

R-sq.(adj) =  0.387   Deviance explained = 39.2%
```

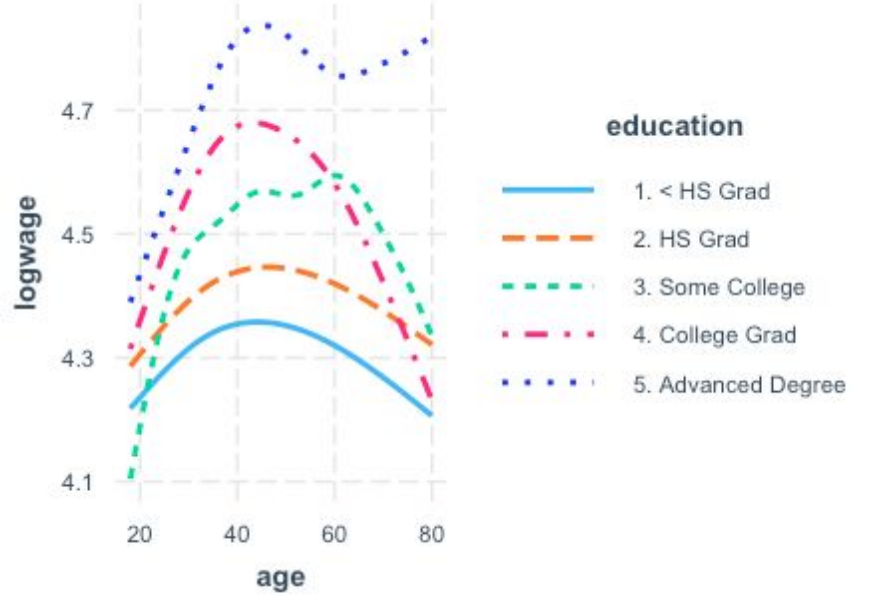
Multivariate GAM

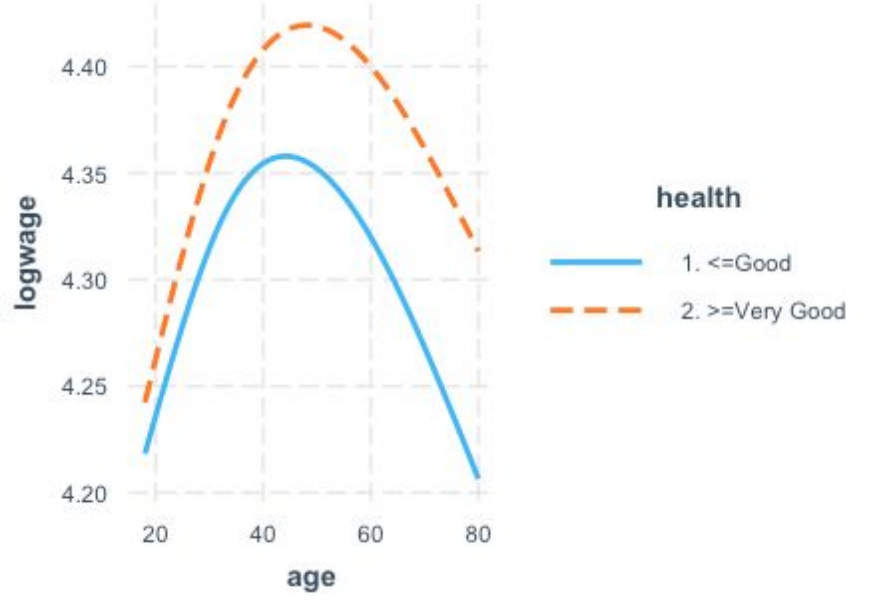
From the plot we can see the transformation of the first model (with age) and the second model (with age^2) is exactly the same. This is because that GAM model can automatically find out the best smoothing line to fit the model. Hence, some data transformations that are necessary in traditional linear regression models are not necessary in GAMs, which is one of the advantages GAMs have over linear models.



Adding Factor-Smooth Interactions

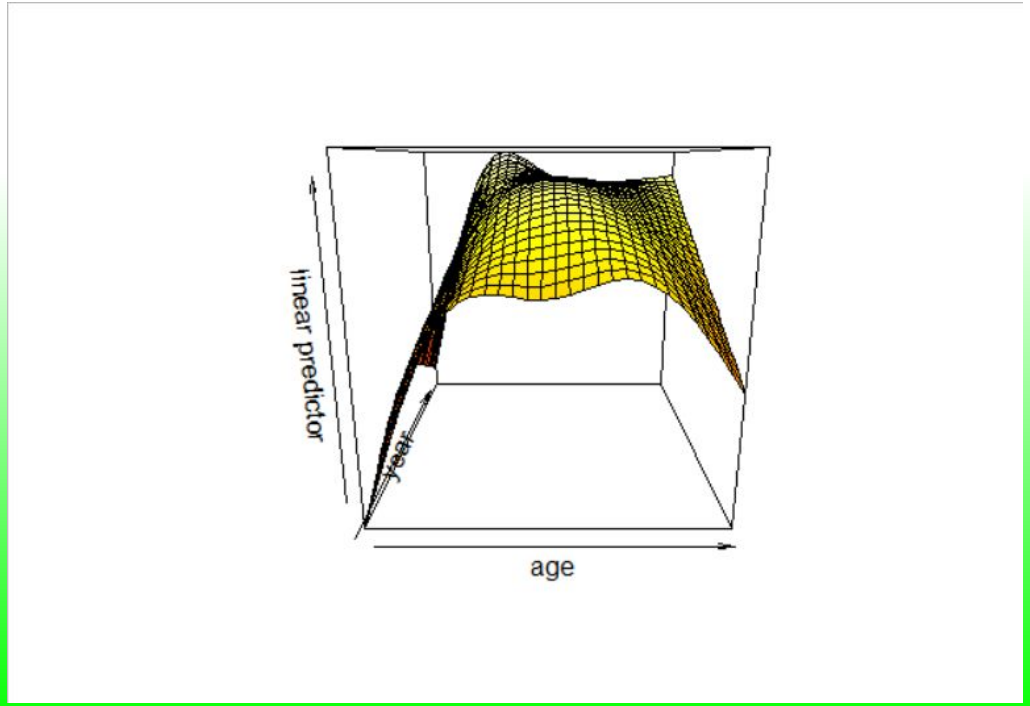
Previous results show that education class and health condition may have a significant effect. Therefore, we assume that wage may have different effects for people of different education classes and health conditions. Therefore, we fit different smooths of wage for people of different education classes and health conditions.





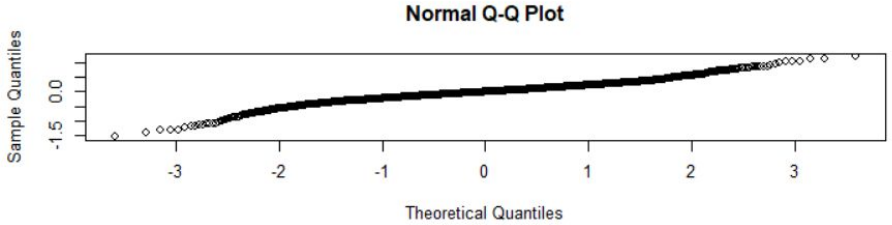
Adding Tensor Interactions

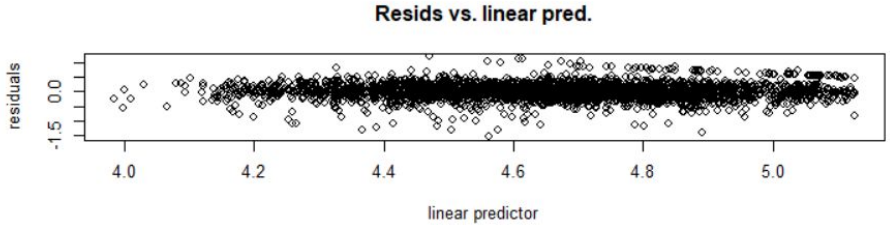
This figure shows that age and year may have complex surfaces. As they have different scales, we add a tensor interaction to our model.

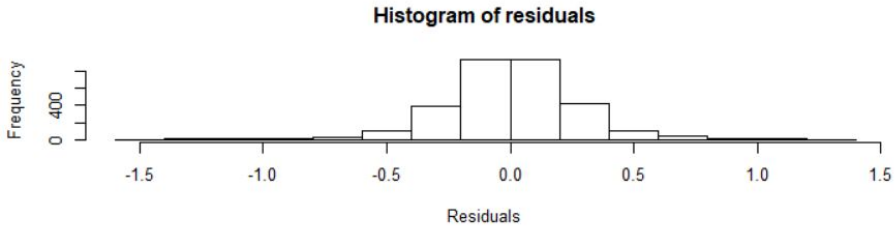


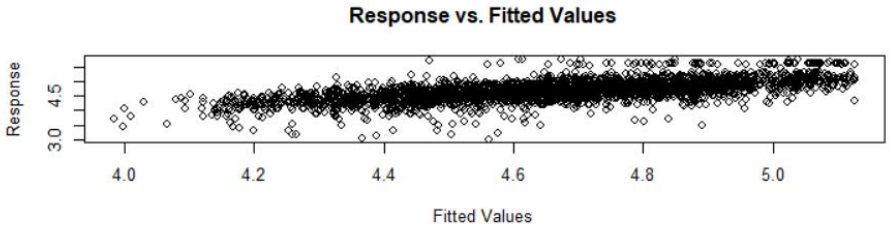
Multivariate GAM

As a result, the adjusted R-squared increases to 0.397 and the deviance explained increases to 40.4%. From the plots shown below, we can find the the Q-Q plot is near a straight line, which means the residuals are normally distributed. The histogram of residuals also shows that the residuals are bell-shaped. The residuals vs linear prediction and the residuals vs fitted values also seem to be linear, which is not bad.









R Output

```
logwage ~ year + s(age, by = education) + maritl + education +
    jobclass + health_ins + age:health + te(age, year)

Parametric coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -5.134e-04  3.913e-04  -1.312   0.1897
year                     2.200e-03  3.476e-05  63.306   <2e-16 ***
maritl2. Married         1.240e-01  1.460e-02   8.497   <2e-16 ***
maritl3. Widowed         3.377e-02  6.475e-02   0.521   0.6021
maritl4. Divorced        4.102e-03  2.370e-02   0.173   0.8626
maritl5. Separated       7.831e-02  3.932e-02   1.992   0.0465 *
education.L              3.262e-01  1.613e-02  20.226   <2e-16 ***
education.Q              2.523e-02  1.411e-02   1.788   0.0739 .
education.C              9.702e-03  1.133e-02   0.857   0.3917
education^4              9.383e-03  1.054e-02   0.890   0.3735
jobclass2. Information   2.066e-02  1.062e-02   1.945   0.0519 .
health_ins               1.831e-01  1.143e-02  16.022   <2e-16 ***
age:health1. <=Good     -5.549e-04  1.638e-03  -0.339   0.7348
age:health2. >=Very Good 7.799e-04  1.649e-03   0.473   0.6362
```

```
Approximate significance of smooth terms:
                                    edf Ref.df     F  p-value
s(age):education2. HS Grad        1.000  1.000 0.246  0.61992
s(age):education3. Some College   5.328  6.413 2.796  0.01202 *
s(age):education4. College Grad   2.977  3.745 3.917  0.00404 **
s(age):education5. Advanced Degree 3.284  4.092 2.853  0.02236 *
te(age,year)                     10.399 13.098 3.604 1.02e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Rank: 72/74
R-sq.(adj) =  0.397   Deviance explained = 40.4%
GCV = 0.075549  Scale est. = 0.074642  n = 3000
```

Model Selection

Model selection is simple to implement for GAMs with only smooth parameters. We can penalize “completely smooth” (non-wiggly) functions so that if their smoothing parameters tend to infinity, the term will be selected out of the model. This is known as null space penalization.

The difficulty is in implementing model selection when dealing with categorical data. We have to consider all possible factor-smooth interactions, as well as factors that are significant to our model but don’t interact with our smooths.

To build a parsimonious GAM, we recommend only including factor-smooth interactions if the interaction has a large effect on the model, and to use backward selection by p-value (removing highest p-value predictor and refitting until all predictors are significant) to remove extraneous predictors.

Model Selection

Using our previous model with null space penalization, we find that the the interaction term of age and health is not significant, so we remove it. We're left with only significant predictors. Notice that the R-squared and deviance explained drop a bit. We're okay with that because we favor a parsimonious model.

```
logwage ~ year + s(age, by = education) + maritl + education +
    jobclass + health_ins + te(age, year)

Parametric coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -10.066150   9.644383  -1.044   0.2967
year                  0.007224   0.004808   1.503   0.1331
maritl2. Married      0.128675   0.014484   8.884   <2e-16 ***
maritl3. Widowed      0.034840   0.064834   0.537   0.5911
maritl4. Divorced     0.003222   0.023676   0.136   0.8918
maritl5. Separated    0.083121   0.039451   2.107   0.0352 *
education.L           0.337291   0.016423  20.538   <2e-16 ***
education.Q           0.027181   0.014437   1.883   0.0598 .
education.C           0.008934   0.011491   0.777   0.4369
education^4           0.009845   0.010599   0.929   0.3530
jobclass2. Information 0.021642  0.010649   2.032   0.0422 *
health_ins            0.186342   0.011442  16.286   <2e-16 ***
```

```
Approximate significance of smooth terms:
                                        edf Ref.df     F  p-value
s(age):education2. HS Grad         3.273e-06      9 0.000 0.888587
s(age):education3. Some College    4.913e+00      9 2.538 7.43e-05 ***
s(age):education4. College Grad    3.007e+00      9 1.893 0.000159 ***
s(age):education5. Advanced Degree 4.084e+00      9 1.680 0.001918 **
te(age,year)                       8.269e+00     23 1.505 3.20e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.392   Deviance explained = 39.8%
GCV = 0.076089  Scale est. = 0.07527   n = 3000
```

Concurvity

Similar to multicollinearity, concurvity is when a smooth term can be approximated by one or more smooth terms in a model. Since our factor-smooth interaction and tensor interaction are both on the age term, we expect to see some degree of concurvity. When checking pairwise concurvity between the smooth terms we are lucky not to see high levels of concurvity, so our model should be adequate.

Limitations and Pitfalls of GAMs

- Factors can’t be smoothed.
- Model selection is tedious and requires strong assumptions.
- Smooth predictors can appear as a “black box.”
- Overfitting.

Resources:

- GAM: The Predictive Modeling Silver Bullet
- Generalized Additive Models in R · A Free Interactive Course
- Thin Plate Spline Regression
- mgcv package
- gam.selection function