Ve406 Lecture 16

Jing Liu

UM-SJTU Joint Institute

July 11, 2018

- Recall the regression spline takes the following form

$$y_i = \hat{g}(x_i) + \hat{e}_i$$

where $\hat{g}$ is a piecewise polynomial that is determined by minimising

$$\sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{g}(x_i))^2$$

which is essentially an extension of SLR to model non-linear relationships

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{e}_i$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are determined by minimising

$$\sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

- It is often used when the relationship between x and $\mathbb{E}[Y \mid X = x]$ is unclear

- Using regression spline save us from aimless trying various transformations on X when the linearity assumption is violated under the simple model

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{e}_i$$

Q: How can we do something similar for ltiple linear regression?

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots \hat{\beta}_k x_{ik} + \hat{e}_i$$

- It is natural to propose the following

$$y_i = \hat{g}_1(x_{i1}) + \hat{g}_2(x_{i2}) + \cdots \hat{g}_k(x_{ik}) + \hat{e}_i$$

where $\hat{g}_j$ is a piecewise polynomial that is determined by minimising

$$\sum_{i=1}^{n} \hat{e}_i^2$$

- However, there is a problem without additional restrictions.

- Consider the simple case where $k = 2$, i.e. the conditional mean is given by

$$\mathbb{E}\left[Y_i \mid X_{i1} = x_{i1}, X_{i2} = x_{i2}\right] = g(x_{i1}) + g(x_{i2})$$

- Now imagine we add a constant c to g_1 and subtract a constant c from g_2,

$$g_1(x_{i1}) + c \qquad \text{for all} \quad x_{i1}$$

$$g_2(x_{i1}) - c \qquad \text{for all} \quad x_{i2}$$

  then nothing observable has changed about the model,

$$\mathbb{E}\left[Y_i \mid X_{i1} = x_{i1}, X_{i2} = x_{i2}\right] = g(x_{i1}) + g(x_{i2})$$

  this is known as a non-identifiable model in statistics.

- It is similar to when we have perfect multicollinearity in MLR, i.e. singular

$$\mathbf{X}^{\mathrm{T}}\mathbf{X}$$

- Practically, it just means the optimisation procedure for the following will fail

$$\min \sum_{i=1}^{n} \hat{e}_i^2$$

since there are infinitely many solutions.

- The non-identifiable part of model can be eliminated by further restrictions.
- The standard convention in this case is to require

$$\sum_{i=1}^{n} g_j(x_{ij}) = 0 \qquad \text{for all} \quad j$$

and adding a constant to the conditional mean independent of any g_j

$$Y_i = \beta_0 + g_1(x_{i1}) + g_2(x_{i2}) + \cdots g_k(x_{ik}) + \varepsilon_i$$

which then is known as additive model.

- To illustrate additive model, consider the following dataset

`prestige`	Pineo-Porter prestige score for occupation
`income`	Average income
`education`	Average number of years of education

```r
> library(carData) # The dataset is a part of it
> attach(Prestige) # Variables become global
```

```r
> sapply(list(prestige, income, education), summary)
```

```
             [,1]      [,2]     [,3]
Min.     14.80000   611.000  6.38000
1st Qu.  35.22500  4106.000  8.44500
Median   43.60000  5930.500 10.54000
Mean     46.83333  6797.902 10.73804
3rd Qu.  59.27500  8187.250 12.64750
Max.     87.20000 25879.000 15.97000
```

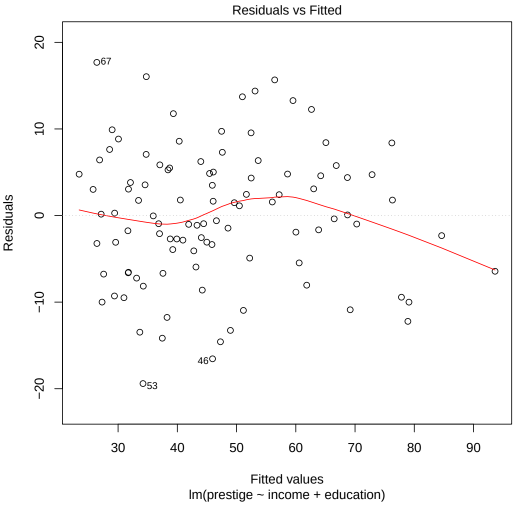
```r
> pre.LM = lm(prestige~income+education)
```

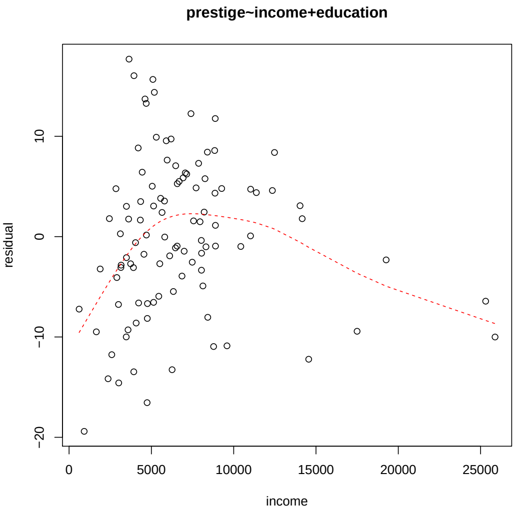
```
> summary(pre.LM)
```

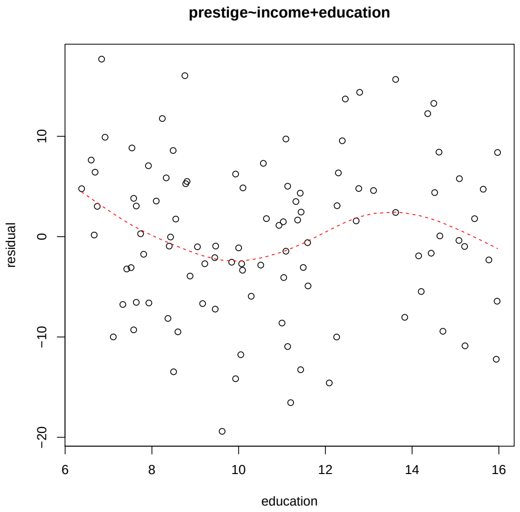
```
Call:
lm(formula = prestige ~ income + education)

Residuals:
     Min       1Q   Median       3Q      Max
-19.4040  -5.3308   0.0154   4.9803  17.6889

Coefficients:
              Estimate Std. Error t value Pr(>t)
(Intercept) -6.8477787  3.2189771  -2.127   0.0359 *
income       0.0013612  0.0002242   6.071 2.36e-08 ***
education    4.1374444  0.3489120  11.858  < 2e-16 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 7.81 on 99 degrees of freedom
Multiple R-squared:  0.798,     Adjusted R-squared:  0.7939
F-statistic: 195.6 on 2 and 99 DF,  p-value: < 2.2e-16
```

- Running the diagnostics shows we might have non-linearity issue,

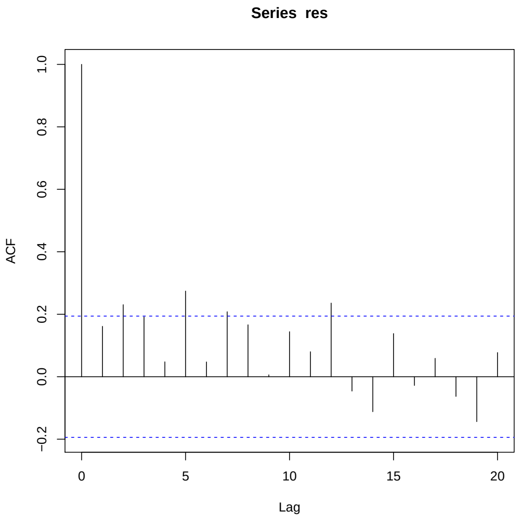
```
> plot(pre.LM, which = 1)
```


Residuals vs Fitted
Residuals
20
10
0
−10
−20
67
46
53
30
40
50
60
70
80
90
Fitted values
lm(prestige ~ income + education)

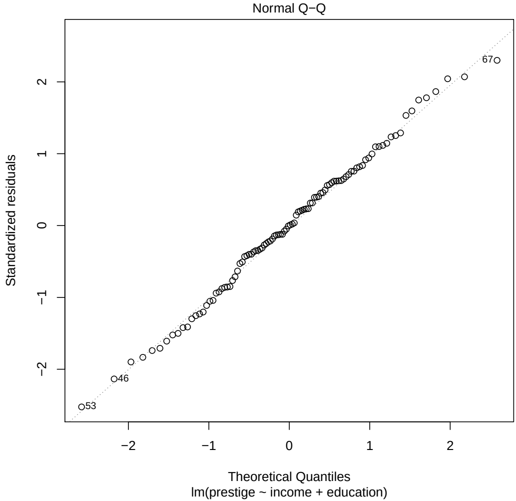

prestige~income+education
residual
income
10
0
−10
−20
0
5000
10000
15000
20000
25000


prestige~income+education
residual
education
10
0
−10
−20
6
8
10
12
14
16

- And the errors might lack independence as well



- However, normality seems to be OK



- So we will try to first linearity instead of transforming the response, and wow we can do so using try regression spline instead of polynomial regression.

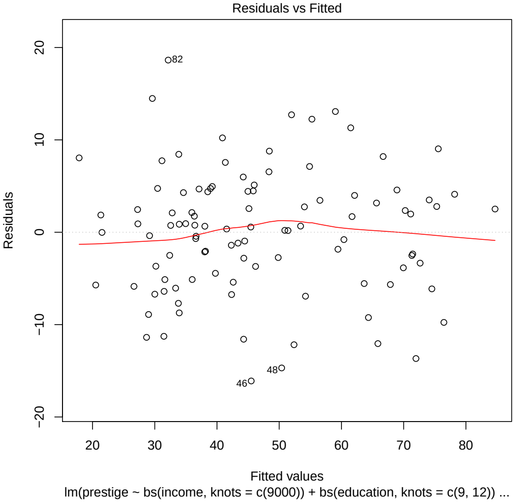
```
> pre.spline.LM =
+   lm(prestige~bs(income, knots = c(9000))
+      +bs(education, knots = c(9, 12)))
```

  where the choices of knots are made largely based on residual plots and the knowledge regarding education system, your guess is just as good as mine.

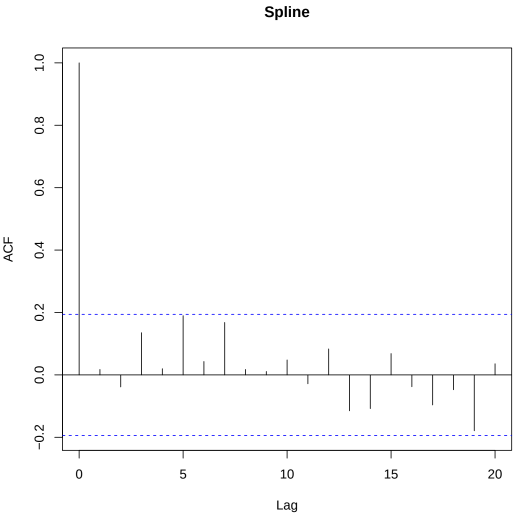
- Looking at diagnostics again

```
> plot(pre.spline.LM, which = 1)
>
> plot(pre.spline.LM, which = 2)
>
> res = pre.spline.LM$residuals
>
> acf(res)
```

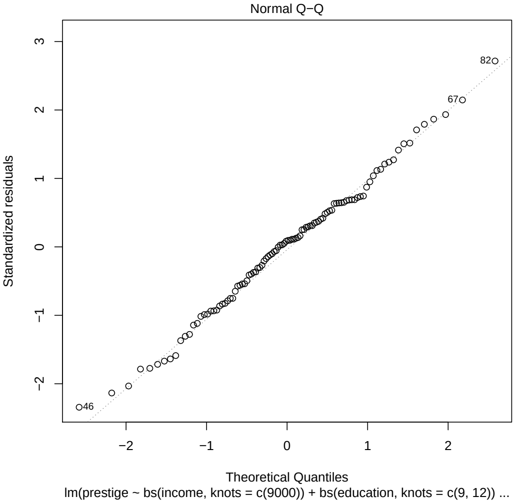
- It seems that having one knot for `income` and two knots for `education` is enough to fix the non-linearity problem.



- It seems also alleviate the independence problem



- And normality seems to be fine for this model as well



- Additive models keep a lot of the nice properties of linear model,

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots \hat{\beta}_k x_{ik} + \hat{e}_i$$

  but the simple interpretation of $\hat{\beta}_j$ is no longer there

$$\beta_j = \frac{\partial}{\partial x_{ij}} \mathbb{E}\left[Y_i \mid \mathbf{X}\right]$$

- The change in $\mathbb{E}\left[Y_i \mid \mathbf{X}\right]$ as x_{ij} changes depending on the value of x_{ij}

$$y_i = \hat{\beta}_0 + \hat{g}_1(x_{i1}) + \hat{g}_2(x_{i2}) + \cdots \hat{g}_k(x_{ik}) + \hat{e}_i$$

- For given x_{ij} value, while holding other predictors constants,

$$g_j$$

  plays the same role as β_j, thus g_j is known as partial response function.

- This means instead of interpreting a constant

$$\hat{\beta}_j$$

when studying the effect of x_{ij} on the conditional mean,

$$\mathbb{E}\left[Y_i \mid \mathbf{X}\right]$$

we have to look at the whole picture.

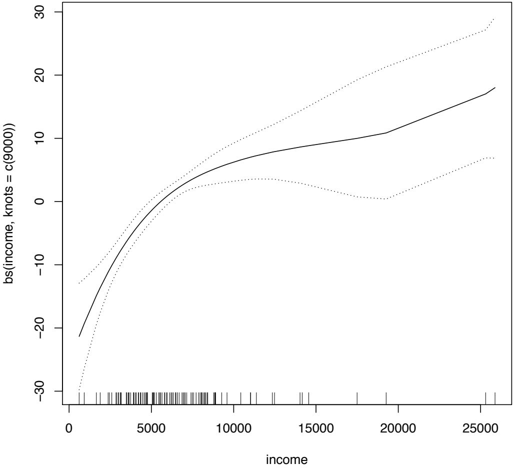
- Instead of manually producing the graph of

$$\hat{g}_j$$

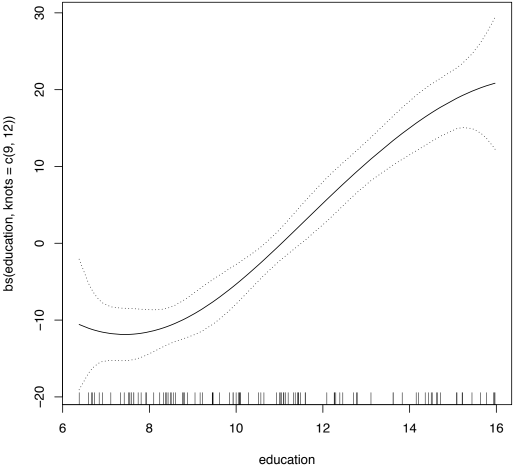
we can use the following function which comes with a confidence band

```
> library(gam)
> plot.Gam(pre.spline.LM, se = TRUE)
```

for understanding and presenting the partial effect of x_{ij} has on $\mathbb{E}\left[Y_i \mid \mathbf{X}\right]$.



bs(income, knots = c(9000))
30
20
10
0
−10
−20
−30
0
5000
10000
15000
20000
25000
income


bs(education, knots = c(9, 12))
30
20
10
0
−10
−20
6
8
10
12
14
16
education

- So far we have only considered additive models that use splines, that is,

$$g_j(x_{ij})$$

is modelled by a piecewise polynomial of x_{ij}.

- In fact, any other type of functions can be used in additive models as long as

$$Y_i = \mathbb{E}\left[Y_i \mid \mathbf{X}\right] + \varepsilon_i$$

where the conditional mean is modelled by

$$\mathbb{E}\left[Y_i \mid \mathbf{X}\right] = \beta_0 + g_1(x_{i1}) + g_2(x_{i2}) + \cdots + g_k(x_i k)$$

- Thus technically your polynomial regression model is also additive model.
- And using smoothing splines and a linear function are certainly allowed.

- To illustrate such additive model, consider the following dataset again

`wage`	Raw wage in the Mid-Atlantic region	
`age`	Age of the worker	
`year`	The year that wage information was recorded	
`education`	A factor with levels:	1. $<$ HS Grad 2. HS Grad 3. Some College 4. College Grad 5. Advanced Degree

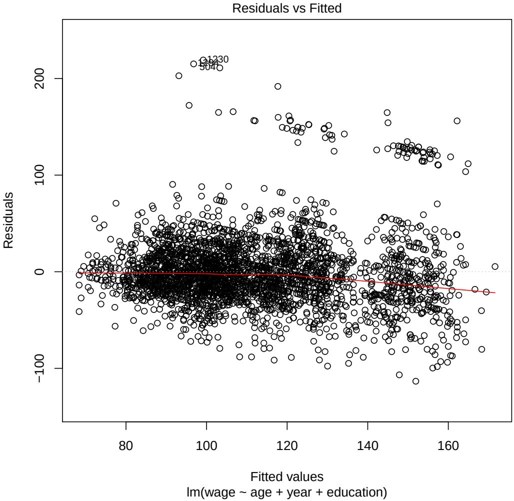
```
> library(ISLR) # The Wage dataset is a part of it
> attach(Wage) # Variables in Wage become global
>
> wage.LM = lm(wage~age+year+education)
```

```
> summary(wage.LM)
```

```
Call:
lm(formula = wage ~ age + year + education)

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 ***
year                         1.056e+00  3.238e-01   3.262  0.00112 **
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 ***
---
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
```

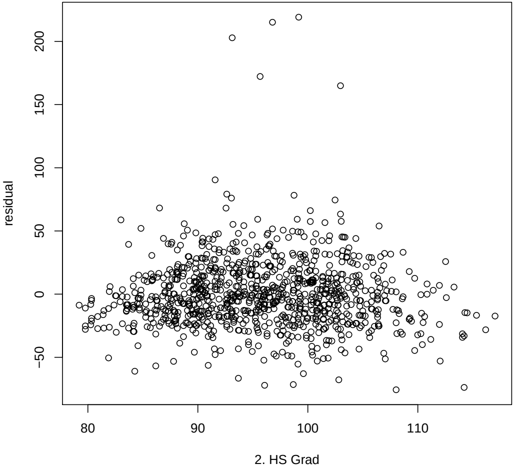
- The model seems to be very promising, but running the diagnostics, we got a rather ugly residual plot, which suggests we should investigate further.

- Initially, I thought it must be due to `education`.




wage~age+year+education
residual
60
40
20
0
−20
−40
−60
70
80
90
100
1. < HS Grad

wage~age+year+education

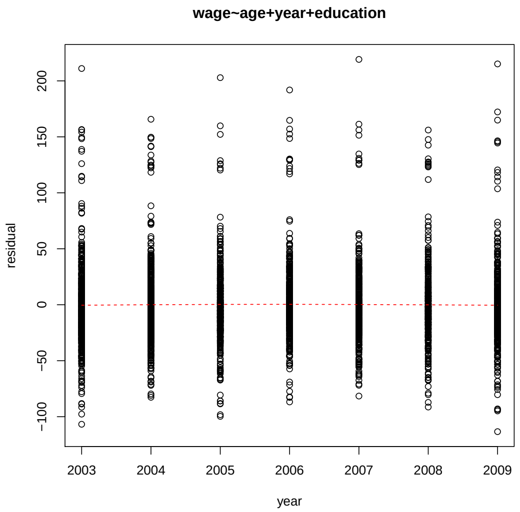



wage~age+year+education
residual
200
150
100
50
0
−50
−100
95
100
105
110
115
120
125
3. Some College

wage~age+year+education
residual
150
100
50
0
−50
−100
110
120
130
140
4. College Grad

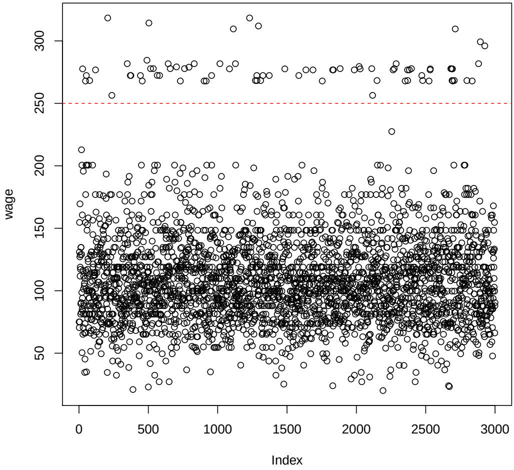

wage~age+year+education
residual
150
100
50
0
−50
−100
140
145
150
155
160
165
170
5. Advanced Degree


wage~age+year+education
residual
200
150
100
50
0
−50
−100
20
30
40
50
60
70
80
age



- It seems there is no variable inside the dataset can be used to explain the group of usually hight wages. So we have to treat them as outliers.

```r
> plot(wage); abline(h=250, col = "red", lty = 2)
```



- Splitting the data, and investigate them individually,

```
> wage.l250.df =
+   subset(Wage, wage<250,
+          select = c(wage, age, year, education))
>
> attributes(wage.l250.df)$row.names =
+   1:nrow(wage.l250.df)
>
> wage.g250.df =
+   subset(Wage, wage>250,
+          select = c(wage, age, year, education))
>
> attributes(wage.g250.df)$row.names =
+   1:nrow(wage.g250.df)
```

- In practice, we do that to compare the differences between the two portions of the dataset, and in the hope that more information regarding the dataset become available in the future, and allows to explain the difference.

- Fit the linear model again,

```
> wage.1250.LM = lm(wage~age+year+education,
+                   data = wage.1250.df)
```

it seems the model assumptions are reasonably good except normality.

```
> shapiro.test(wage.1250.LM$residuals)
```

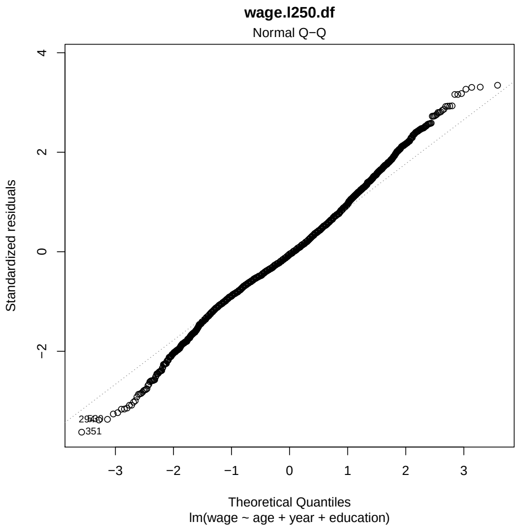
```
        Shapiro-Wilk normality test

data:  wage.1250.LM$residuals
W = 0.99288, p-value = 8.495e-11
```

- Because the dataset is still reasonably large,

```
> nrow(wage.1250.df)
```

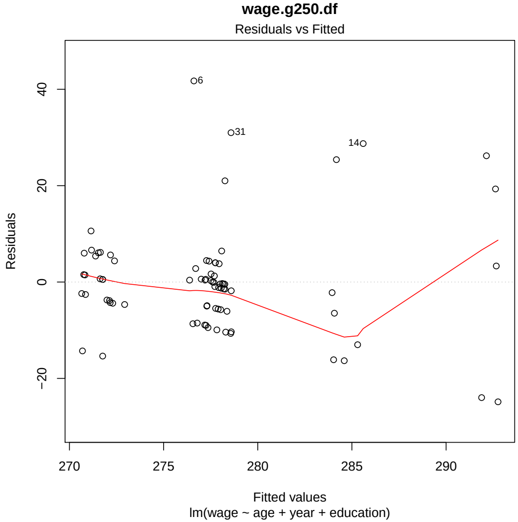
```
[1] 2921
```

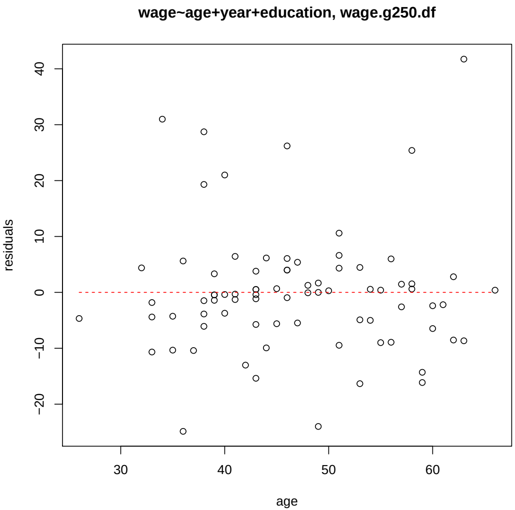
and QQ-normal plot reveals the distribution is fairly symmetric, so we will reply on the central limit theorem and not consider further transformation.

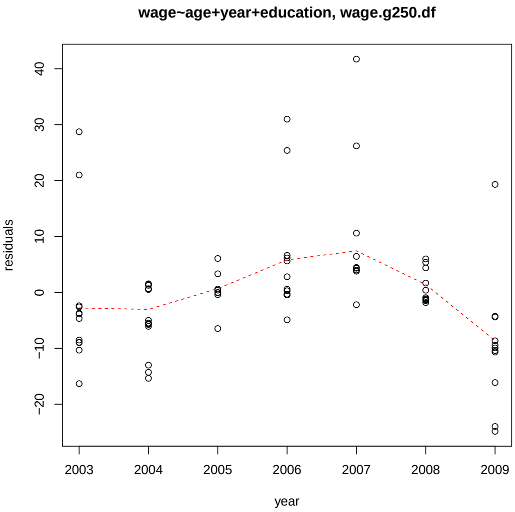

wage.l250.df
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(wage ~ age + year + education)
351
−3
−2
−1
0
1
2
3
4

- For the other portion, additive model with smoothing spline is worth trying

```
> wage.g250.LM = lm(wage~age+year+education,
+                  data = wage.g250.df)
```




wage~age+year+education, wage.g250.df
residuals
40
30
20
10
0
−10
−20
30
40
50
60
age


wage~age+year+education, wage.g250.df
residuals
year
40
30
20
10
0
−10
−20
2003
2004
2005
2006
2007
2008
2009

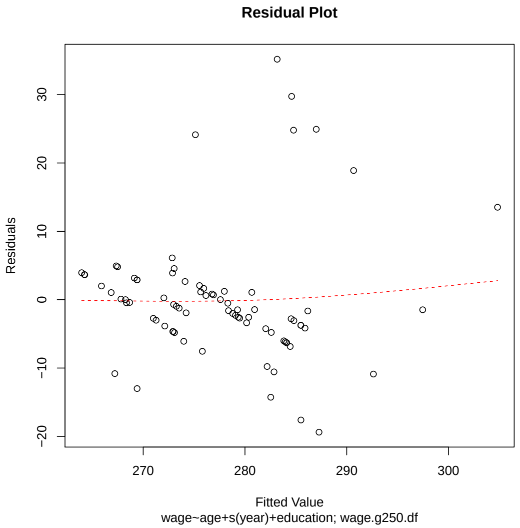
- It seems the following additive model is reasonable but far from perfect

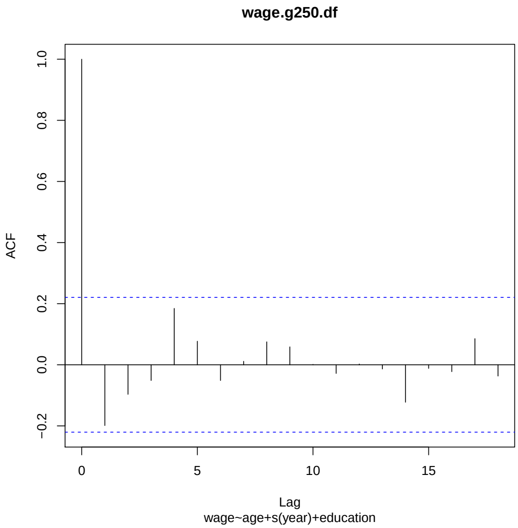
```
> library(gam)
> wage.g250.SMS = gam(wage~age+s(year)+education,
+                          data = wage.g250.df)
```

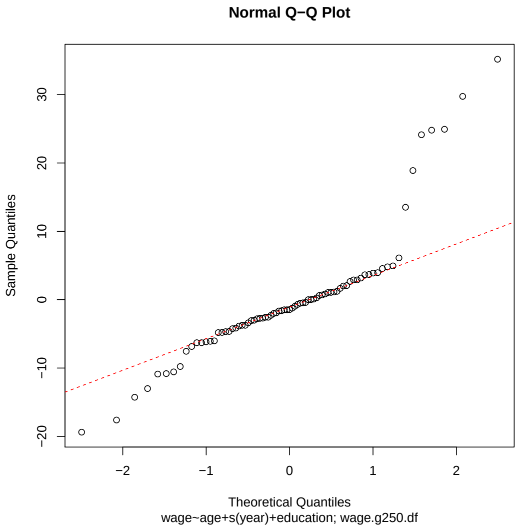
- The Residual plot shows the linearity assumption is OK, but the amount of variability in `year` is not great, and the variance might not be equal.
- Together with the ACF plot, it seems the independence assumption is fine.
- It it clear that normality is violated, but I could not fix it after trying a few common transformations. I decided to move on which means we should not use any results that are based on normality, that includes t-test, etc.

```
> shapiro.test(wage.g250.SMS$residuals)
```

```
        Shapiro-Wilk normality test

data:  wage.g250.SMS$residuals
W = 0.82966, p-value = 4.131e-08
```


Residual Plot
Residuals
Fitted Value
wage~age+s(year)+education; wage.g250.df
30
20
10
0
−10
−20
270
280
290
300


wage.g250.df
ACF
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
0
5
10
15
Lag
wage~age+s(year)+education


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
wage~age+s(year)+education; wage.g250.df
30
20
10
0
−10
−20
−2
−1
0
1
2

```
> summary(wage.g250.SMS)
```

```
Call: gam(formula = wage ~ age + s(year) + education, data = wage.g250.df)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-19.378   -4.198   -1.445    2.034   35.176

(Dispersion Parameter for gaussian family taken to be 96.4459)

    Null Deviance: 11930.06 on 78 degrees of freedom
Residual Deviance: 6751.217 on 70.0001 degrees of freedom
AIC: 595.5866

Number of Local Scoring Iterations: 2

Anova for Parametric Effects
          Df Sum Sq Mean Sq F value    Pr(>F)
age        1  110.6  110.59  1.1467    0.2879
s(year)    1  100.3  100.35  1.0404    0.3112
education  3 2641.1  880.37  9.1282 3.545e-05 ***
Residuals 70 6751.2   96.45
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Anova for Nonparametric Effects
            Npar Df Npar F     Pr(F)
(Intercept)
age
s(year)           3 10.085 1.324e-05 ***
education
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
```