Wage Example – Polynomial Regression

Part 1: Polynomial Regression

Wage Example – Polynomial Regression

Basic Implementation

First, load the required package and read the wage table.

```
#load the required packages
library(ISLR)
```

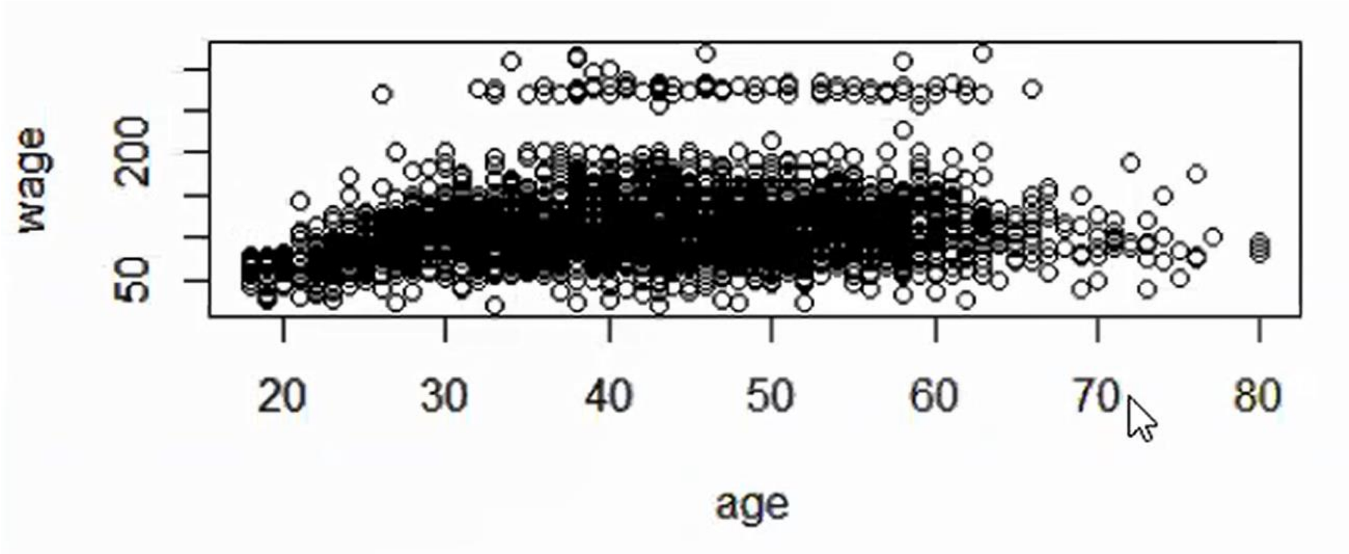
```
attach(Wage)
```

Wage Example – Polynomial Regression

Basic Implementation

Plot wage against age

```
plot(age,wage)
```



Wage Example – Polynomial Regression

Basic Implementation

You can run a simple linear regression with age, age^2 , age^3 and age^4 and then print a summary of the regression

```
#fit a regression line
fitla<-lm(wage~age+I(age^2)+I(age^3)+I(age^4),data=Wage
)
#print a summary of the regression
summary(fitla)
```

Note: must have indicator function on age^2 , age^3 and age^4, if not, output will only have one coefficient on age

Wage Example – Polynomial Regression

Basic Implementation

With indicator function, the printed summary of the regression is as follows:

```
Call:
lm(formula = wage ~ age + I(age^2) + I(age^3) + I(age^4), data = Wage)

Residuals:
    Min      1Q  Median      3Q     Max 
-98.707 -24.626  -4.993  15.217 203.693 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.842e+02  6.004e+01  -3.067 0.002180 ** 
age          2.125e+01  5.887e+00   3.609 0.000312 ***
I(age^2)    -5.639e-01  2.061e-01  -2.736 0.006261 ** 
I(age^3)     6.811e-03  3.066e-03   2.221 0.026398 *  
I(age^4)    -3.204e-05  1.641e-05  -1.952 0.051039 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 39.91 on 2995 degrees of freedom
Multiple R-squared:  0.08626,	Adjusted R-squared:  0.08504 
F-statistic: 70.69 on 4 and 2995 DF,  p-value: < 2.2e-16
```

Note the coefficients are: -184.2, 21.25, -0.564, etc.

Wage Example – Polynomial Regression

Basic Implementation

No indicator function on age^2 , age^3 and age^4 : do NOT work

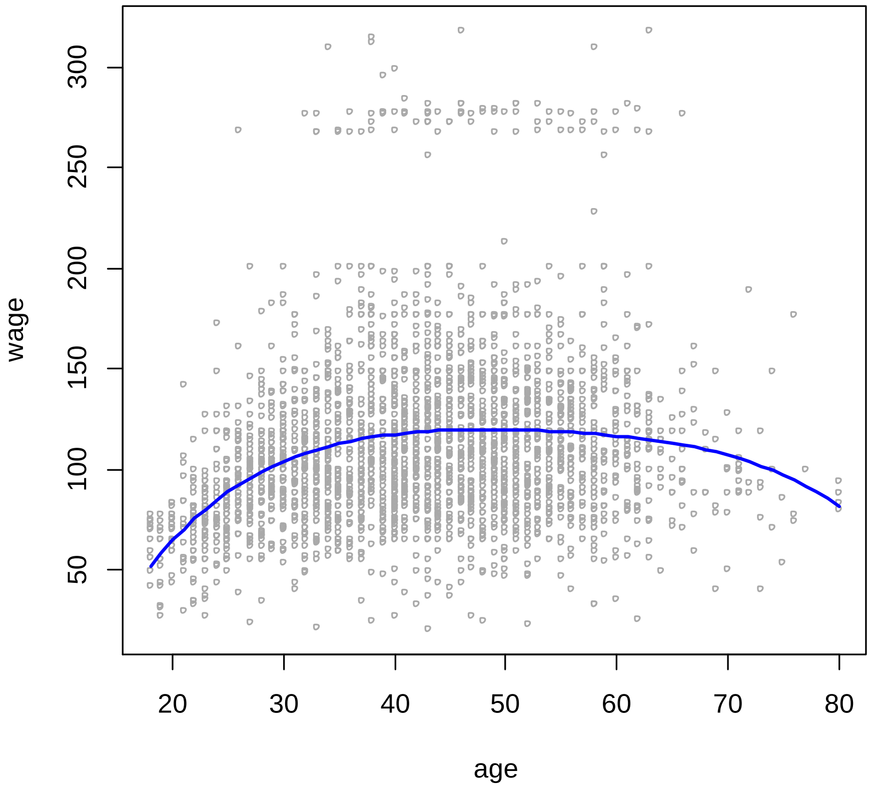
```
# as comparison
fit2 <- lm(wage~age+age^2+age^3+age^4,data = Wage)
summary(fit2)

Call:
lm(formula = wage ~ age + age^2 + age^3 + age^4, data = Wage)

Residuals:
     Min       1Q   Median       3Q      Max 
-100.265  -25.115   -6.063   16.601  205.748 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 81.70474    2.84624   28.71   <2e-16 ***
age          0.70728    0.06475   10.92   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 40.93 on 2998 degrees of freedom
Multiple R-squared:  0.03827,	Adjusted R-squared:  0.03795 
F-statistic: 119.3 on 1 and 2998 DF,  p-value: < 2.2e-16
```

Wage Example – Polynomial Regression

Basic Implementation

In order to visualize the fit, we can plot a fitted line for different ages.

```
#Set up age grid for prediction
agelims=range(age)
age.grid=seq(from=agelims[1],to=agelims[2])
#Predict the grid
preds<-predict(fitla,newdata=list(age=age.grid),se=TRUE)
plot(age,wage,xlim=agelims,cex=.5,col="darkgray")
lines(age.grid,preds$fit,lwd=2,col="blue")
```

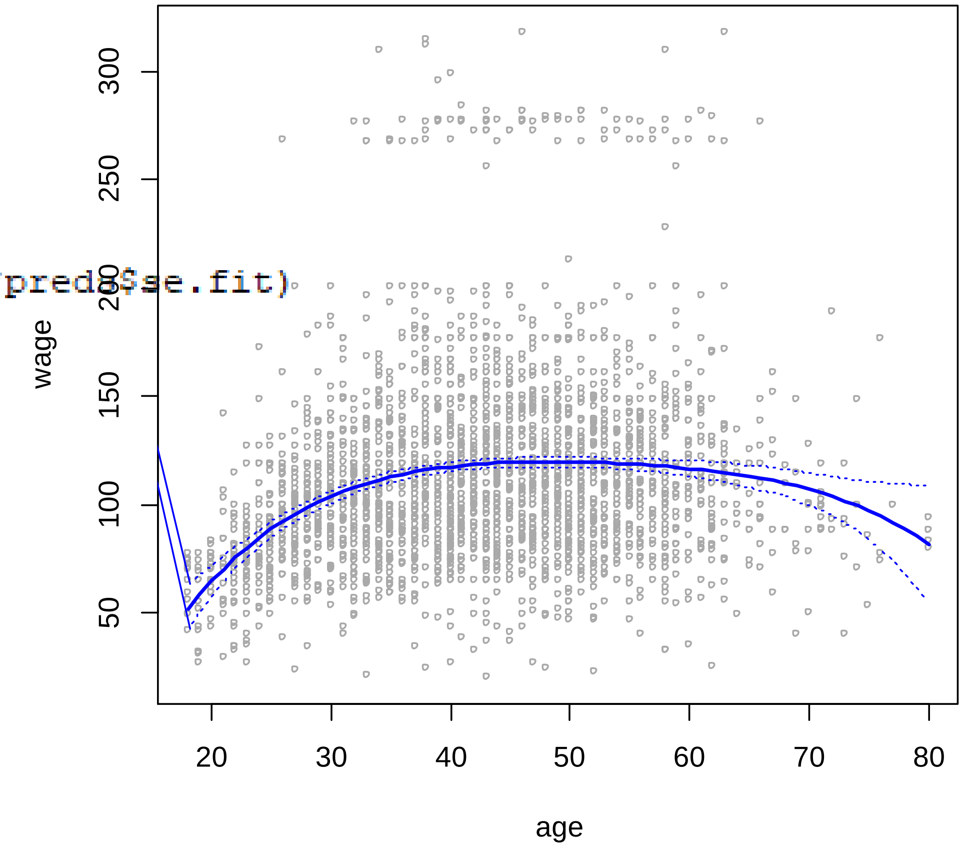


Wage Example – Polynomial Regression

Basic Implementation

We may wish to plot the confidence intervals for different ages.

```
#construct confidence interval
se.bands <-cbind(preds$fit+2*preds$se.fit,preds$fit-2*preds$se.fit)
#plot confidence interval with the existing graph
matlines(age.grid,se.bands,lwd=1, col='blue',lty=3)
```



Wage Example – Polynomial Regression

Basic Implementation

We can also shorten age, age^2 , age^3 and age^4 with poly function.

raw = T vs raw = F (default)

```
fit3<-lm(wage~poly(age,4,raw=T),data=Wage)
summary(fit3)
Call:
lm(formula = wage ~ poly(age, 4, raw = T), data = Wage)

Residuals:
    Min      1Q  Median      3Q     Max
-98.707 -24.626  -4.993  15.217 203.693

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -1.842e+02  6.004e+01  -3.067 0.002180 **
poly(age, 4, raw = T)1   2.125e+01  5.887e+00   3.609 0.000312 ***
poly(age, 4, raw = T)2  -5.639e-01  2.061e-01  -2.736 0.006261 **
poly(age, 4, raw = T)3   6.811e-03  3.066e-03   2.221 0.026398 *
poly(age, 4, raw = T)4  -3.204e-05  1.641e-05  -1.952 0.051039 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 39.91 on 2995 degrees of freedom
Multiple R-squared:  0.08626,	Adjusted R-squared:  0.08504
F-statistic: 70.69 on 4 and 2995 DF,  p-value: < 2.2e-16
```

Reason: Numeric stability

When raw = T, coefficients can be directly interpreted, but numerically unstable (larger no. raised to a large power)

When raw = F, coefficients cannot be directly interpreted, but numerically stable (R uses some linear algebra technique)

If raw=T is numerical stable, both raw=T/F are mathematically equivalent & give the same prediction, so in practice should use raw = F (default)

Wage Example – Polynomial Regression

Basic Implementation

The estimates, standard errors and t values becomes different when raw is not set to true.

```
fit4<-lm(wage~poly(age,4),data=Wage)
summary(fit4)
```

```
Call:
lm(formula = wage ~ poly(age, 4), data = Wage)

Residuals:
    Min      1Q  Median      3Q     Max 
-98.707 -24.626  -4.993  15.217 203.693 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)    111.7036     0.7287 153.283  < 2e-16 ***
poly(age, 4)1  447.0679    39.9148  11.201  < 2e-16 ***
poly(age, 4)2 -478.3158    39.9148 -11.983  < 2e-16 ***
poly(age, 4)3  125.5217    39.9148   3.145  0.00168 ** 
poly(age, 4)4  -77.9112    39.9148  -1.952  0.05104 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 39.91 on 2995 degrees of freedom
Multiple R-squared:  0.08626,	Adjusted R-squared:  0.08504 
F-statistic: 70.69 on 4 and 2995 DF,  p-value: < 2.2e-16
```

Default: raw = F

coeff. different from raw = T

Wage Example – Polynomial Regression

Basic Implementation

But raw=T & raw=F are mathematically equivalent

```
preds_rawT<-predict(fit3,newdata=list(age=age.grid),se=TRUE)
preds_rawF<-predict(fit4,newdata=list(age=age.grid),se=TRUE)
preds_rawT$fit[1:5]
preds_rawF$fit[1:5]
```

```
> preds_rawT$fit[1:5]
       1        2        3        4        5 
51.93145 58.49674 64.57188 70.18273 75.35440 
> preds_rawF$fit[1:5]
       1        2        3        4        5 
51.93145 58.49674 64.57188 70.18273 75.35440 
```

Wage Example – Polynomial Regression

Basic Implementation

How many polynomial term we should include in the model? What is the criteria of evaluating different models?

We can use ANOVA to compare models with polynomial degree 1, 2, 3, 4 and 5

```
#fit five polynomial regression models
fit.1<-lm(wage~age,data=Wage)
fit.2<-lm(wage~poly(age,2),data=Wage)
fit.3<-lm(wage~poly(age,3),data=Wage)
fit.4<-lm(wage~poly(age,4),data=Wage)
fit.5<-lm(wage~poly(age,5),data=Wage)
#Use anova to evaluate the five models
anova(fit.1,fit.2,fit.3,fit.4,fit.5)
```

Recall anova in 7002
H_0: Simpler model sufficient vs H_a: More complex model required
To use anova(), models must be nested (i.e. model's predictors subset of another)

Wage Example – Polynomial Regression

Basic Implementation

Output of Anova:

```
Analysis of Variance Table

Model 1: wage ~ age
Model 2: wage ~ poly(age, 2)
Model 3: wage ~ poly(age, 3)
Model 4: wage ~ poly(age, 4)
Model 5: wage ~ poly(age, 5)
  Res.Df     RSS Df Sum of Sq        F    Pr(>F)
1   2998 5022216
2   2997 4793430  1    228786 143.5931 < 2.2e-16 ***
3   2996 4777674  1     15756   9.8888  0.001679 **
4   2995 4771604  1      6070   3.8098  0.051046 .
5   2994 4770322  1      1283   0.8050  0.369682
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What conclusion can we draw from this table?

Wage Example – Polynomial Regression

Basic Implementation

Output of Anova:

```
Analysis of Variance Table

Model 1: wage ~ age
Model 2: wage ~ poly(age, 2)
Model 3: wage ~ poly(age, 3)
Model 4: wage ~ poly(age, 4)
Model 5: wage ~ poly(age, 5)
  Res.Df     RSS Df Sum of Sq        F    Pr(>F)
1   2998 5022216
2   2997 4793430  1    228786 143.5931 < 2.2e-16 ***
3   2996 4777674  1     15756   9.8888  0.001679 **
4   2995 4771604  1      6070   3.8098  0.051046 .
5   2994 4770322  1      1283   0.8050  0.369682
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What conclusion can we draw from this table?

p-value stat. sig.: model 2 and 1 are sig. diff., indicating quad term necessary

Similarly model 3 sig. diff. from 2, but model 4 NOT sig. diff. from 3.

Should choose the model 3 (poly. deg. 3), based on ANOVA.

Wage Example – Polynomial Regression

Basic Implementation

We can include more variables (e.g. education), some without transformation, some with transformation (poly)

Note: education is categorical (no transformation needed), coded as dummy

```
#fit five polynomial regression models
fit.1a<-lm(wage~education+age,data=Wage)
fit.2a<-lm(wage~education+poly(age,2),data=Wage)
fit.3a<-lm(wage~education+poly(age,3),data=Wage)
fit.4a<-lm(wage~education+poly(age,4),data=Wage)
fit.5a<-lm(wage~education+poly(age,5),data=Wage)
#Use anova to evaluate the five models
anova(fit.1a,fit.2a,fit.3a,fit.4a,fit.5a)
```

Wage Example – Polynomial Regression

Basic Implementation

The printed ANOVA table is as follows:

```
Analysis of Variance Table

Model 1: wage ~ education + age
Model 2: wage ~ education + poly(age, 2)
Model 3: wage ~ education + poly(age, 3)
Model 4: wage ~ education + poly(age, 4)
Model 5: wage ~ education + poly(age, 5)
  Res.Df     RSS Df Sum of Sq        F Pr(>F)
1   2994 3867992
2   2993 3725395  1    142597 114.7077 <2e-16 ***
3   2992 3719809  1      5587   4.4940 0.0341 *
4   2991 3719777  1        32   0.0255 0.8731
5   2990 3716972  1      2805   2.2562 0.1332
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Should choose the model 3 (poly. deg. 3), based on ANOVA.

Wage Example – Polynomial Regression

Basic Implementation

The output for final model is as follows:

```
Call:
lm(formula = wage ~ education + poly(age, 3), data = Wage)

Residuals:
     Min       1Q   Median       3Q      Max 
-114.880  -19.937   -2.967   14.623  214.683 

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                   85.606      2.157  39.693  < 2e-16 ***
education2. HS Grad           10.861      2.434   4.462 8.41e-06 ***
education3. Some College      23.218      2.562   9.064  < 2e-16 ***
education4. College Grad      37.930      2.547  14.894  < 2e-16 ***
education5. Advanced Degree   62.613      2.764  22.655  < 2e-16 ***
poly(age, 3)1                362.668     35.466  10.226  < 2e-16 ***
poly(age, 3)2               -379.777     35.429 -10.719  < 2e-16 ***
poly(age, 3)3                 74.849     35.309   2.120   0.0341 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 35.26 on 2992 degrees of freedom
Multiple R-squared:  0.2877,	Adjusted R-squared:  0.286 
F-statistic: 172.6 on 7 and 2992 DF,  p-value: < 2.2e-16
```

education1 (Intercept) = <HS Grad

education2 compared with education1: increase by $10.86K on average

education3: increase by $23.22K, etc

Wage Example – Poly Logistic Regression

Part 2: Poly Logistic Regression

Wage Example – Poly Logistic Regression

Basic Implementation

I(wage>250) is an indicator of whether wage exceeds 250.

We can also fit a polynomial logistic regression as follows:

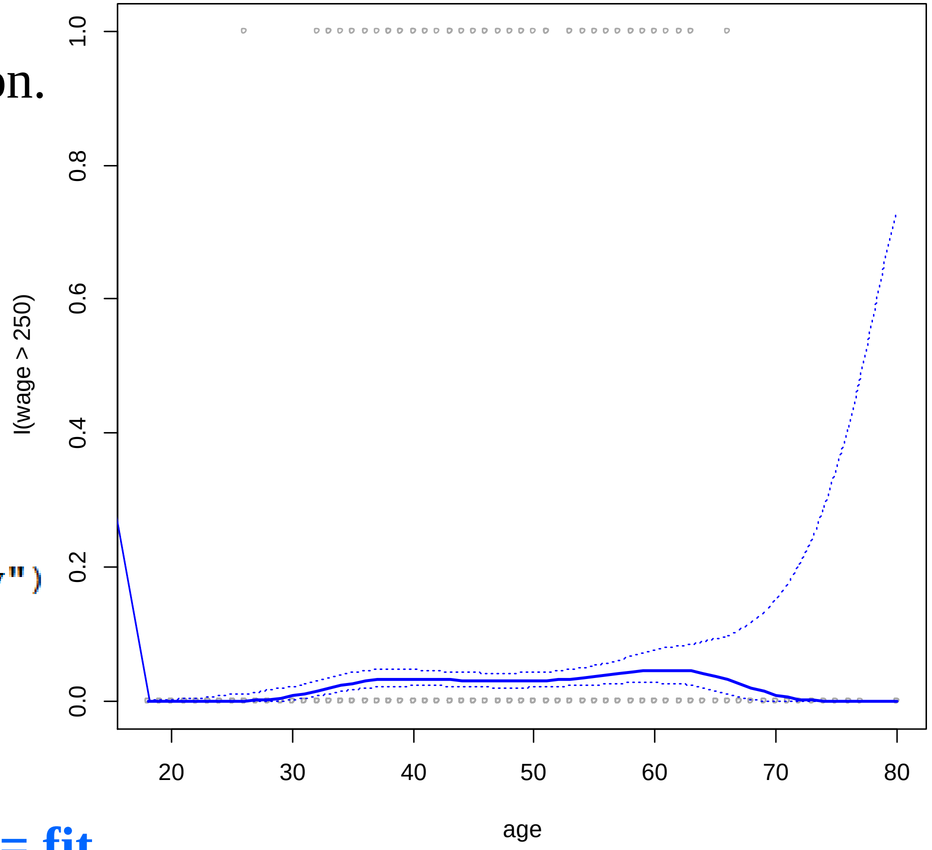
```
#fit a polynomial logistic regression
fit<-glm(I(wage>250)~poly(age,4),data=Wage,family=binomial)
```

Wage Example – Poly Logistic Regression

Basic Implementation

Similarly, we can visualize the output for logistic regression.

```
#Predict the wage for each age
preds<-predict(fit,newdata=list(age=age.grid),se=TRUE)
#Predict the probability for each of age
pfit<-exp(preds$fit)/(exp(preds$fit)+1)
#Construct the confidence interval for each of age
se.bands_logit<-cbind(preds$fit-2*preds$se.fit,
                      preds$fit+2*preds$se.fit)
se.bands<-exp(se.bands_logit)/(1+exp(se.bands_logit))
#Plot a similar plot
plot(age,I(wage>250),xlim=agelims,cex=.5,col="darkgray")
lines(age.grid,pfit,lwd=2,col="blue")
matlines(age.grid,se.bands,lwd=1, col='blue',lty=3)
```



Recall $log\frac{P(G=1)}{P(G=0)} = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 = \text{fit}$

$\Leftrightarrow \frac{P(G=1)}{1-P(G=1)} = e^{fit} \Leftrightarrow P(G=1) = \frac{e^{fit}}{e^{fit}+1}$