

Statistics 452: Statistical Learning and Prediction

Chapter 7, Part 1: Simple Extensions of the Linear Model

Brad McNeney

Polynomial Regression

- ▶ We have used polynomial regression before.
- ▶ This was the standard way to extend the linear model.
- ▶ The text recommends against a degree of more than 4.
- ▶ To avoid collinearity we can use the covariates returned by the `poly()` function (more on these functions in the next set of notes).

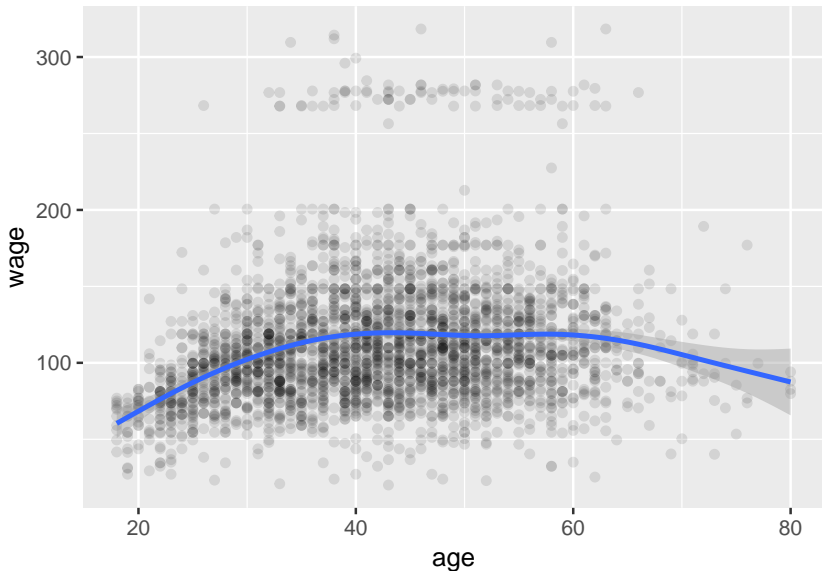
Example: Wage Data

- Predict wages as a function of age.

```
library(ISLR)
data(Wage)
head(Wage,n=3)
```

```
##           year age           maritl      race      education
## 231655 2006   18 1. Never Married 1. White    1. < HS Grad
## 86582 2004   24 1. Never Married 1. White    4. College Grad
## 161300 2003   45           2. Married 1. White 3. Some College
##
##           region      jobclass      health health_ins
## 231655 2. Middle Atlantic 1. Industrial    1. <=Good    2. No
## 86582 2. Middle Atlantic 2. Information 2. >=Very Good 2. No
## 161300 2. Middle Atlantic 1. Industrial    1. <=Good    1. Yes
##
##           logwage      wage
## 231655 4.318063 75.04315
## 86582 4.255273 70.47602
## 161300 4.875061 130.98218
```

```
library(ggplot2)
ggplot(Wage, aes(x=age, y=wage)) + geom_point(alpha=0.1) +
  geom_smooth(formula=formula(wage ~ poly(age, 4)))
```



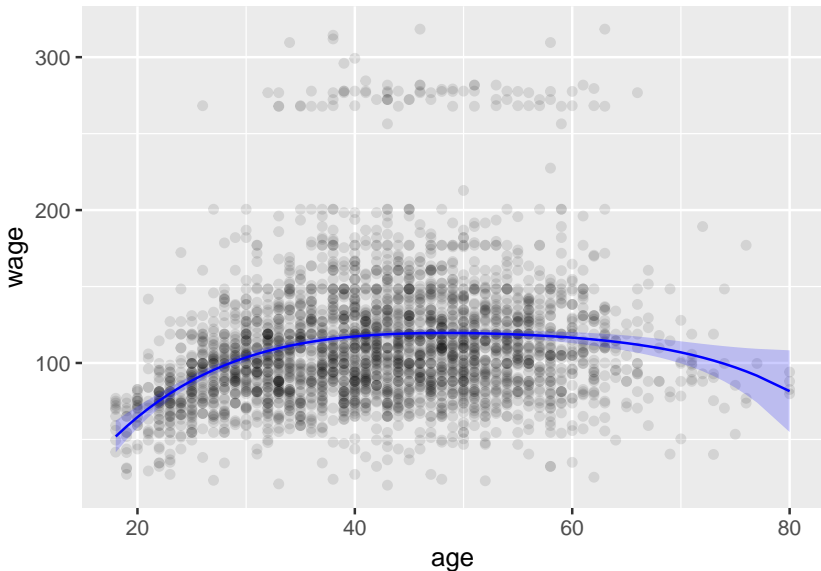
DIY Confidence Bands

- ▶ Use `predict()` to get the predictions and plot CI bands with `geom_ribbon()`
- ▶ First get the predictions and point-wise CIs:

```
wfit <- lm(wage ~ poly(age,4),data=Wage)
newWage <- data.frame(age = sort(unique(Wage$age)))
wpred <- data.frame(newWage,predict(wfit,newdata=newWage,interval="confidence"))
head(wpred,n=3)
```

##	age	fit	lwr	upr
## 1	18	51.93145	41.54284	62.32006
## 2	19	58.49674	49.92674	67.06674
## 3	20	64.57188	57.52864	71.61511

```
ggplot(Wage, aes(x=age, y=wage)) + geom_point(alpha=0.1) +  
  geom_ribbon(aes(x=age, y=fit, ymin=lwr, ymax=upr),  
            data=wpred, fill="blue", alpha=.2) +  
  geom_line(aes(y=fit), data=wpred, color="blue")
```

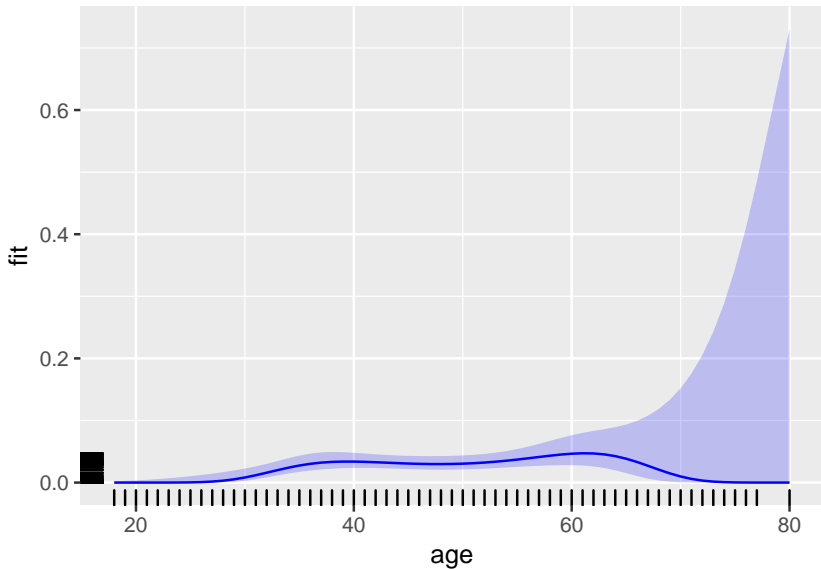


Classification Example: Wage > 250K

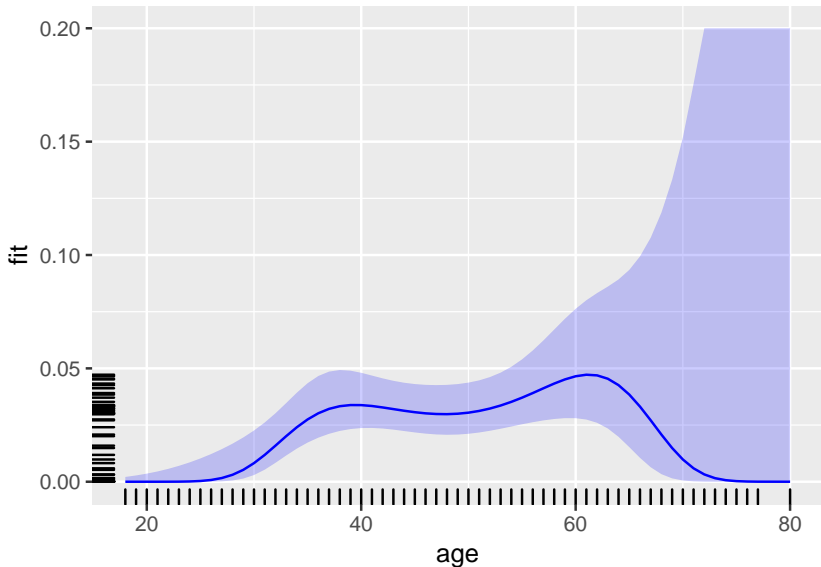
```
library(dplyr)
Wage <- mutate(Wage, wage250 = as.numeric(wage > 250))
wfit <- glm(wage250 ~ poly(age, 4), data=Wage, family=binomial())
wpred <- data.frame(newWage,
  predict(wfit, newdata=newWage, se.fit=TRUE)) # predicted logits
expit <- function(x) { exp(x)/(1+exp(x)) }
wpred <- mutate(wpred,
  lwr=expit(fit-2*se.fit), upr=expit(fit+2*se.fit),
  fit = expit(fit))
# Don't go below 0 or above 1
head(wpred, n=3)
```

##	age	fit	se.fit	residual.scale	lwr	upr
## 1	18	9.826427e-09	6.140231	1	4.560982e-14	0.002112586
## 2	19	7.577844e-08	5.254561	1	2.067720e-12	0.002769461
## 3	20	4.746699e-07	4.464911	1	6.283757e-11	0.003572811

```
ggplot(wpred,aes(x=age,y=fit)) + geom_rug() +  
  geom_ribbon(aes(ymin=lwr,ymax=upr),fill="blue",alpha=.2,limits=c(0,.2)) +  
  geom_line(aes(y=fit),data=wpred,color="blue")
```




```
wpred <- mutate(wpred, upr = pmin(upr, 0.2))  
ggplot(wpred, aes(x=age, y=fit)) + geom_rug() +  
  geom_ribbon(aes(ymin=lwr, ymax=upr), fill="blue", alpha=.2, limits=c(0, .2)) +  
  geom_line(aes(y=fit), data=wpred, color="blue")
```



Step Functions

- ▶ In Epidemiology it is common to discretize age and treat as a categorical variable.
 - ▶ Categorical variables are coded as dummy variables for regression.
- ▶ The regression will fit a separate mean for each category, which is more flexible than, say, linear in age.

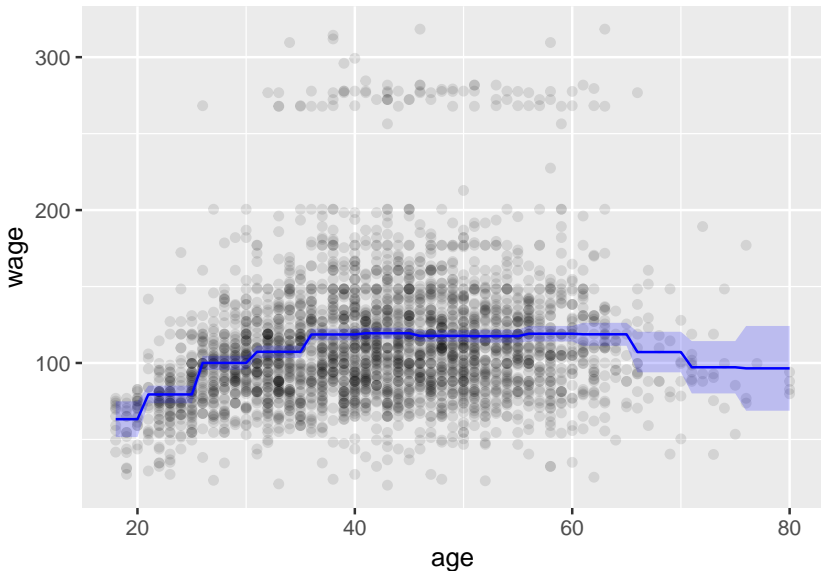
```
with(Wage, range(age))
```

```
## [1] 18 80
```

```
agebreaks <- c(15,20,25,30,35,40,45,50,55,60,65,70,75,80)
wfit <- lm(wage ~ cut(age,breaks=agebreaks),data=Wage)
newWage <- data.frame(age = sort(unique(Wage$age)))
wpred <- data.frame(newWage,predict(wfit,newdata=newWage,interval="confidence"))
head(wpred,n=3)
```

```
##   age    fit    lwr    upr
## 1  18 63.2435 51.55813 74.92886
## 2  19 63.2435 51.55813 74.92886
## 3  20 63.2435 51.55813 74.92886
```

```
ggplot(Wage, aes(x=age, y=wage)) + geom_point(alpha=0.1) +  
  geom_ribbon(aes(x=age, y=fit, ymin=lwr, ymax=upr),  
            data=wpred, fill="blue", alpha=.2) +  
  geom_line(aes(y=fit), data=wpred, color="blue")
```



Classification Example: Wage > 250K

- ▶ No top-earners less than about 25 or older than about 65, so need to re-do the age breaks.

```
library(dplyr)
agebreaks <- c(15,30,35,40,45,50,55,60,80)
wfit <- glm(wage250 ~ cut(age,agebreaks),data=Wage,family=binomial())
wpred <- data.frame(newWage,
  predict(wfit,newdata=newWage,se.fit=TRUE)) # predicted logits
expit <- function(x) { exp(x)/(1+exp(x)) }
wpred <- mutate(wpred,
  lwr=expit(fit-2*se.fit),upr=expit(fit+2*se.fit),
  fit = expit(fit))
head(wpred,n=3)
```

##	age	fit	se.fit	residual.scale	lwr	upr
## 1	18	0.001915711	1.000295	1	0.0002595404	0.01399226
## 2	19	0.001915711	1.000295	1	0.0002595404	0.01399226
## 3	20	0.001915711	1.000295	1	0.0002595404	0.01399226

```
ggplot(wpred,aes(x=age,y=fit)) + geom_rug() +  
  geom_ribbon(aes(ymin=lwr,ymax=upr),fill="blue",alpha=.2,limits=c(0,.2)) +  
  geom_line(aes(y=fit),data=wpred,color="blue")
```

