# Statistics 452: Statistical Learning and Prediction

## Chapter 1: Introduction

Brad Mcneney

2017-09-01

# What is Statistical Learning?

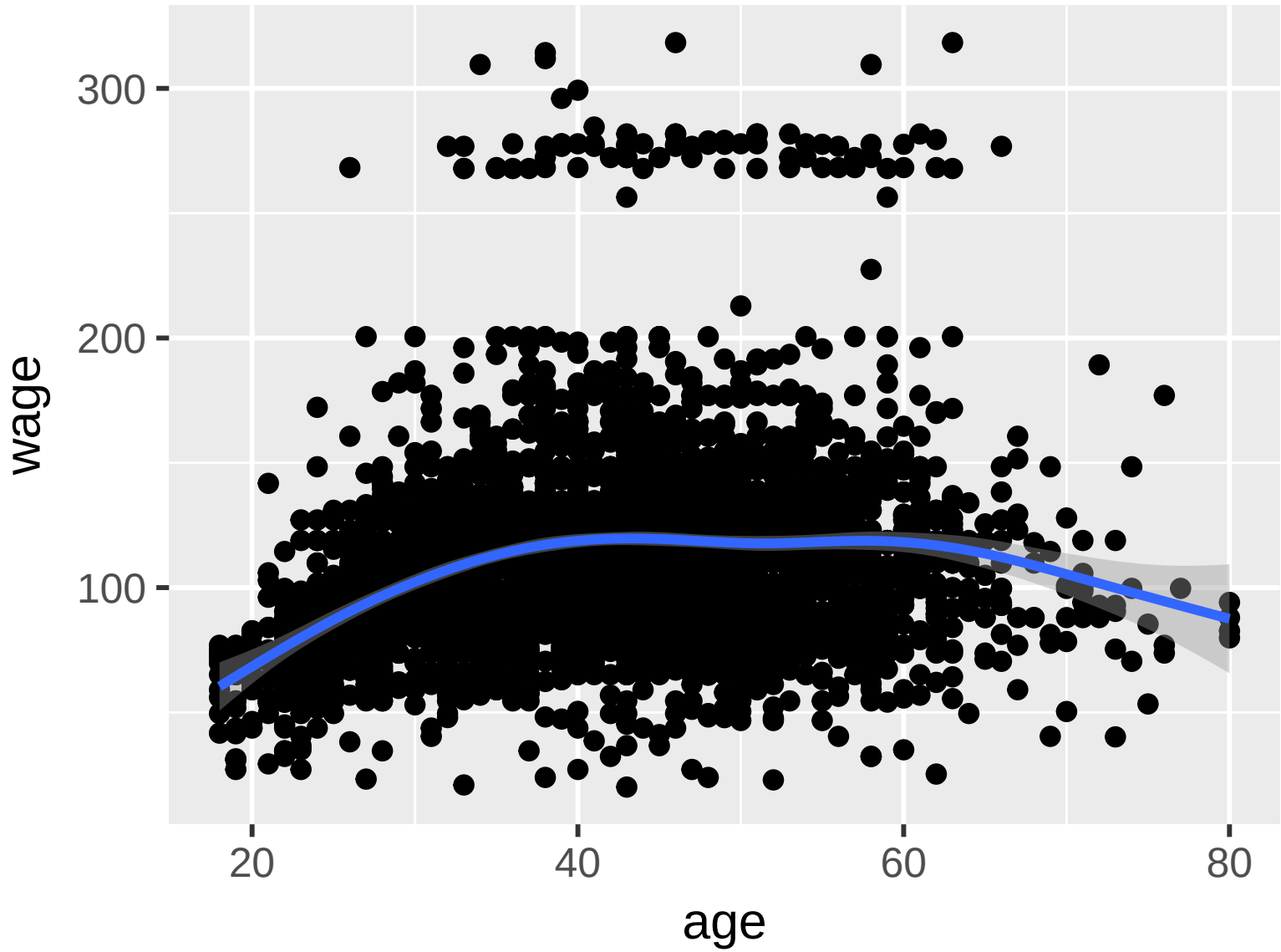
- Tools for learning from data.
- Multiple regression is an example of such a tool:
  - We propose a linear model $Y = \beta_0 + X_1\beta_1 + \ldots X_p\beta_p + \epsilon$, and fit the model.
  - We may interpret fitted coefficients, or use them to obtain predictions.
  - Such a problem is said to be "supervised" because of the response variable $Y$, viewed as an "output" that is influenced by the "inputs" $X_1, \ldots, X_p$.

# Example: The Wage data

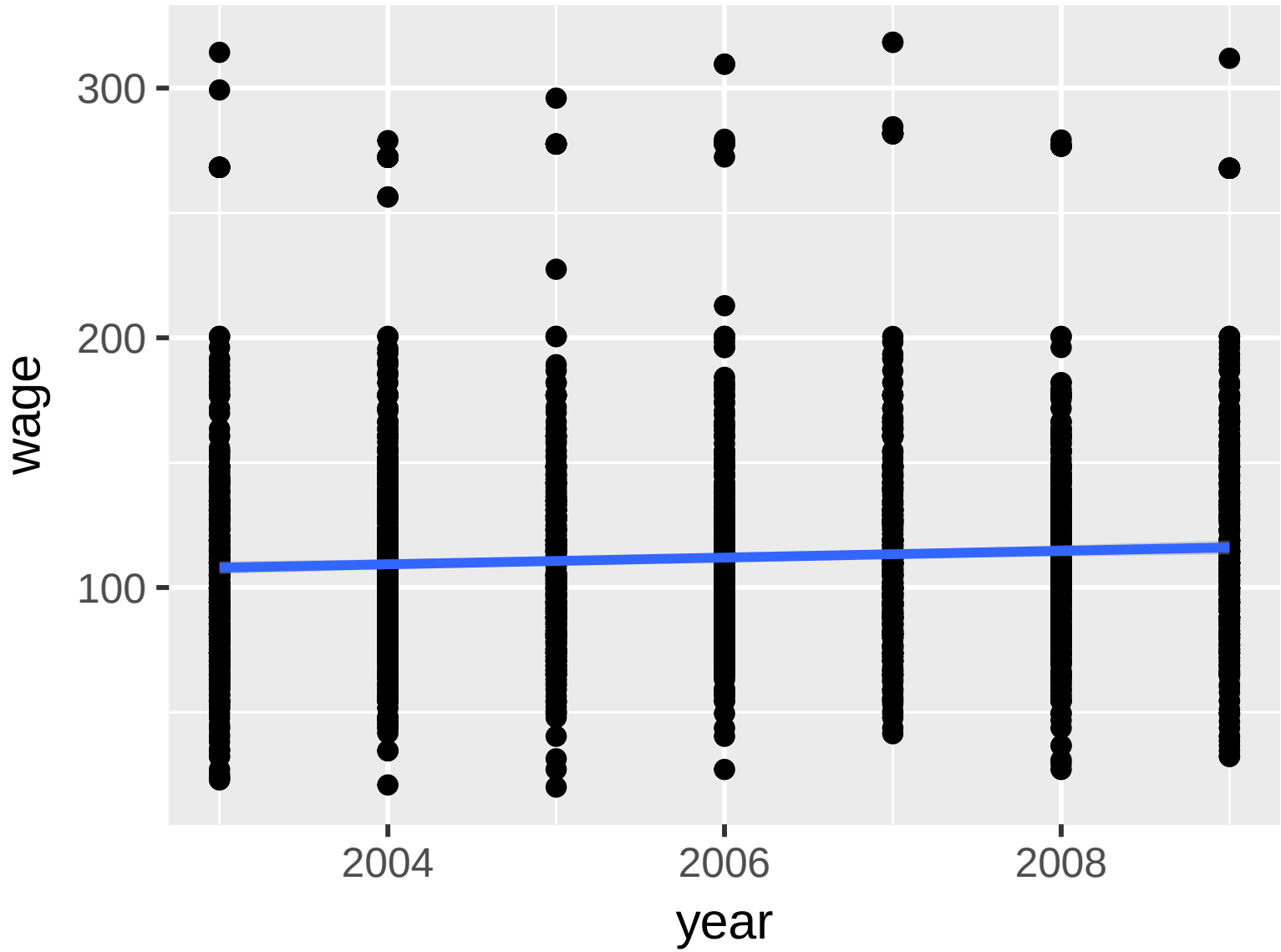
```
library(ISLR)
data(Wage)
head(Wage)
```

```
##        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
## 155159 2003  43       2. Married 3. Asian 4. College Grad
## 11443  2005  50      4. Divorced 1. White      2. HS Grad
## 376662 2008  54       2. Married 1. White 4. College Grad
##                      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
## 155159 2. Middle Atlantic 2. Information 2. >=Very Good     1. Yes
## 11443  2. Middle Atlantic 2. Information      1. <=Good     1. Yes
## 376662 2. Middle Atlantic 2. Information 2. >=Very Good     1. Yes
##         logwage      wage
## 231655 4.318063  75.04315
## 86582  4.255273  70.47602
## 161300 4.875061 130.98218
## 155159 5.041393 154.68529
## 11443  4.318063  75.04315
## 376662 4.845098 127.11574
```

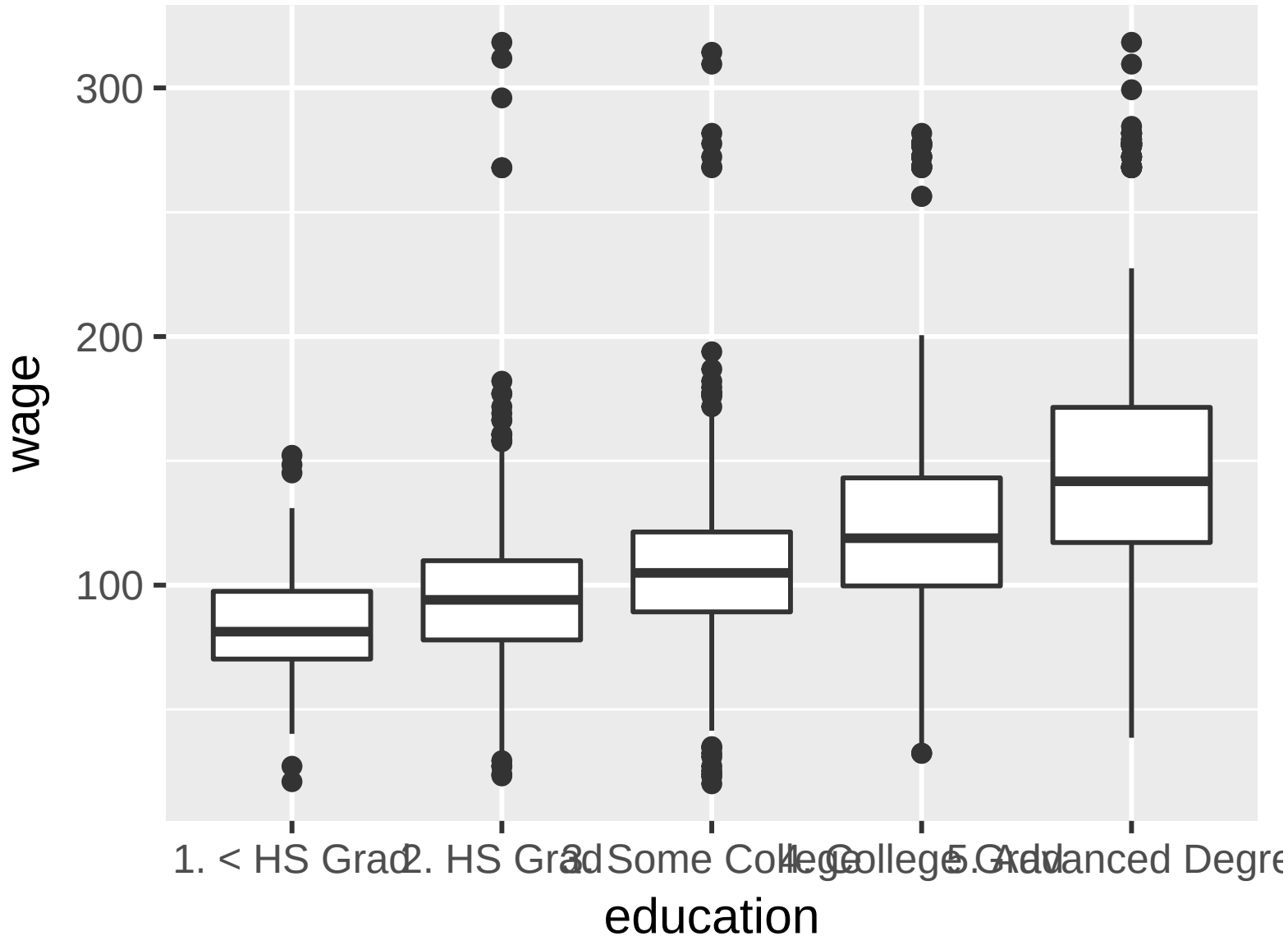
```
library(ggplot2)
ggplot(Wage,aes(x=age,y=wage)) +
  geom_point() + geom_smooth()
```



```
ggplot(Wage,aes(x=year,y=wage)) + geom_point() + geom_smooth(method="lm")
```



```r
library(ggplot2)
ggplot(Wage,aes(x=education,y=wage)) +
  geom_boxplot()
```



```r
wfit <- lm(wage ~ age + I(age^2) + year + education, data=Wage)
summary(wfit)$coefficients
```

```
##                                      Estimate   Std. Error    t value
## (Intercept)                     -2307.7808304 6.375417e+02  -3.619811
## age                                 4.2360813 3.443460e-01  12.301818
## I(age^2)                           -0.0424098 3.921848e-03 -10.813726
## year                                1.1445212 3.177882e-01   3.601522
## education2. HS Grad                10.7519901 2.430424e+00   4.423916
## education3. Some College           23.2956075 2.558074e+00   9.106698
## education4. College Grad           37.9663708 2.542784e+00  14.931023
## education5. Advanced Degree        62.6013504 2.759253e+00  22.687786
##                                       Pr(>|t|)
## (Intercept)                       2.997042e-04
## age                               5.780734e-34
## I(age^2)                          9.200003e-27
## year                              3.215041e-04
## education2. HS Grad               1.003869e-05
## education3. Some College          1.513602e-19
## education4. College Grad          1.122880e-48
## education5. Advanced Degree      2.775558e-105
```

# Other tools

- We will study non-linear methods for supervised learning, and methods appropriate to "unsupervised" problems, where there is no response variable.

# Notation

- There are $n$ distinct observations.
- The random response for the $i$th individual is $Y_i$ and observed value is $y_i$; $i = 1 \ldots, n$.
- There are $p$ explanatory variables $X = (X_1, \ldots, X_p)$.
- The measured value of the $j$th explanatory variable on the $i$th observation is denoted $x_{ij}$

# More Notation and Simple Matrix Algebra

- Will to discuss in tutorial for those who need a refresher or quick intro.

# R and RStudio

- You will need to install both R and the RStudio interface, **and** you will need to create an RStudio "project" based on the class GitHub repository.
  - See the computer software "getting started" page on canvas.
  - Also, Will to discuss in tutorial.
- Also be sure to install the `tidyverse` and `ISLR` packages using the RStudio `Tools` menu, or from the command line

```
install.packages(c("tidyverse","ISLR")
```