Leture 12

Brad McNeney

Preparation for the final exam

- Final exam is from 8:30 - 11:30, Tuesday April 16, in Images Theatre.
- Though the time slot for the final is three hours, I will aim for a final that will take most people two hours.
  - A bit less than twice as long as the midterm.
- The exam is cumulative, but with about 2/3 emphasis on material after the midterm (lectures 7-11) and 1/3 from before (lectures 1-6).
- In cases where we discussed both base-R and tidyverse approaches to a task, you are responsible only for the tidyverse version.
- The exam is closed book. R cheatsheets will be provided.
  - Cheatsheets for this year's exam available at https://canvas.sfu.ca/courses/43617/files/9846163/download?wrap=1

Course objectives – recap

- Understand basic R data structures and programming
- Learn how to use base R and R package functions for data management, exploration, presentation and analysis
- Learn how to use packages from the “tidyverse”, a collection of modern tools for data science.
  - https://www.tidyverse.org/

Overview of lectures

Focus on the following topics from lectures 1-11:

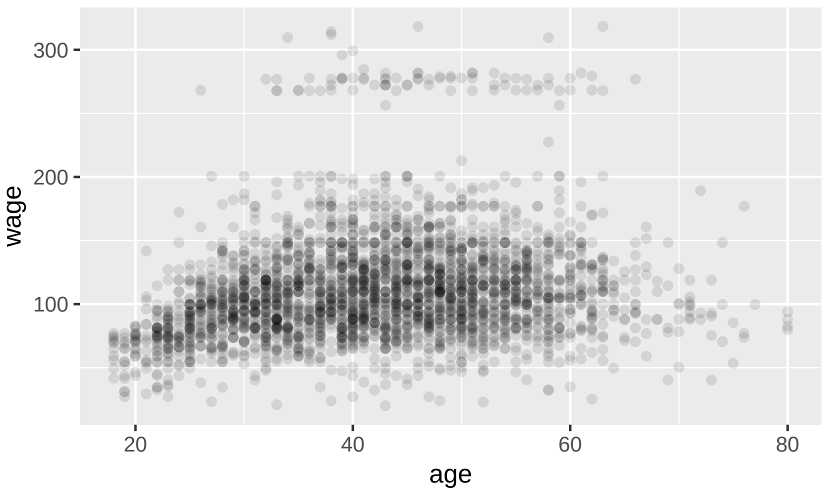
1. (no topics – introduction and getting started)
2. dataframes, lists, vectors, functions
3. subsetting with $, [and [[and with dplyr, for loops, reading data from files
4. transforming variables, working with factors, working with dates
5. working with strings
6. reading from databases, merging/joining database tables and dataframes
7. what is tidy data, reshaping with gather and spread (homework 2), split-apply-combine for transformations and data summaries
8. iterating with map, graphics with ggplot2
9. graphics with ggplot2
10. pseudo-random number generation, permutation tests, the replicate function for simulation
11. the bootstrap, cross-validation

More Examples

- using ggplot, gather, split-apply-combine, map

```r
library(tidyverse)
```

```r
library(ISLR); data(Wage)
ggplot(Wage,aes(x=age,y=wage)) + geom_point(alpha=.1)
```

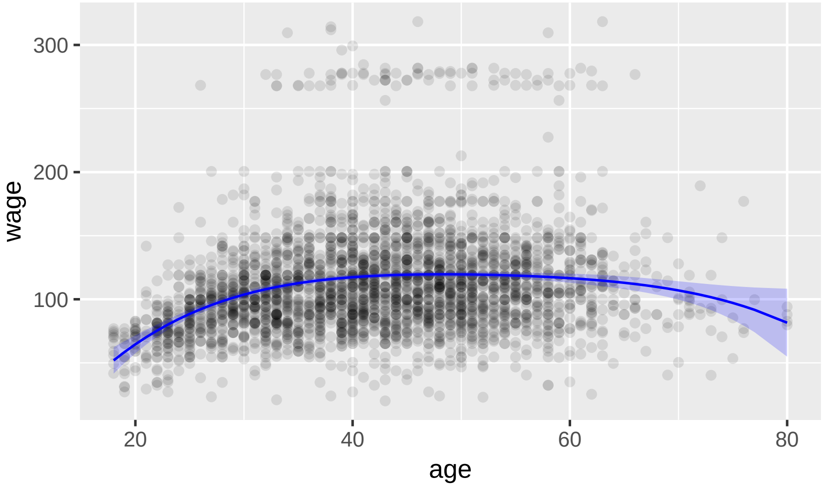


```r
fit <- lm(wage ~ poly(age,4),data=Wage,model=TRUE)
summary(fit)$coef
```

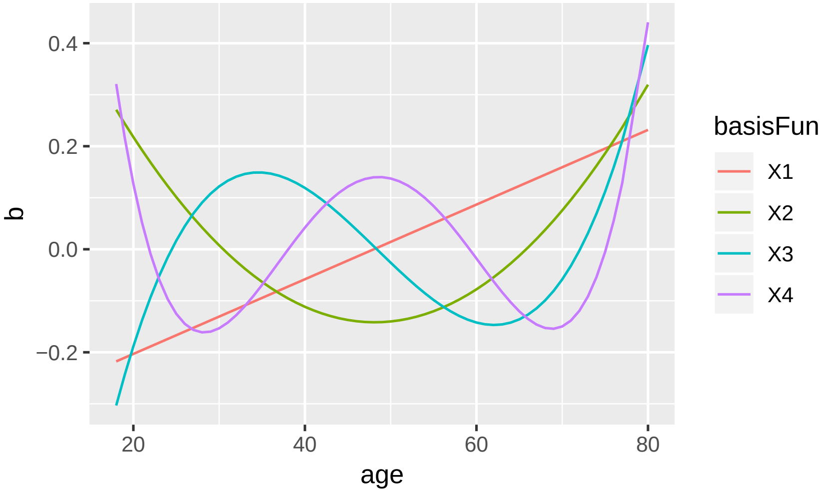
```
##                 Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)    111.70361  0.7287409 153.283015 0.000000e+00
## poly(age, 4)1  447.06785 39.9147851  11.200558 1.484604e-28
## poly(age, 4)2 -478.31581 39.9147851 -11.983424 2.355831e-32
## poly(age, 4)3  125.52169 39.9147851   3.144742 1.678622e-03
## poly(age, 4)4  -77.91118 39.9147851  -1.951938 5.103865e-02
```

```r
plotfit <- function(fit,dat,newdat){
  preds <- data.frame(newdat,
          predict(fit,newdata=newdat,interval="confidence"))
  ggplot(dat,aes(x=age)) + geom_point(aes(y=wage),alpha=0.1) +
    geom_ribbon(aes(ymin=lwr,ymax=upr),
                data=preds,fill="blue",alpha=.2) +
    geom_line(aes(y=fit),data=preds,color="blue")
}
```

```r
age <- sort(unique(Wage$age))
newdat <- data.frame(age=age)
plotfit(fit,Wage,newdat)
```



```r
age <- sort(unique(Wage$age))
Xmat <- data.frame(age=age,poly(age,4))
Xlong <- gather(Xmat,basisFun,b,-age)
ggplot(Xlong,aes(x=age,y=b,color=basisFun)) + geom_line()
```



```r
Wage %>% group_by(maritl) %>% summarize(n=n())
```

```
## # A tibble: 5 x 2
##   maritl                n
##   <fct>             <int>
## 1 1. Never Married    648
## 2 2. Married         2074
## 3 3. Widowed           19
## 4 4. Divorced         204
## 5 5. Separated         55
```

```r
Wage <- mutate(Wage,maritl2 = fct_lump(maritl,n=2))
Wage %>% group_by(maritl2) %>% summarize(n=n())
```

```
## # A tibble: 3 x 2
##   maritl2               n
##   <fct>             <int>
## 1 1. Never Married    648
## 2 2. Married         2074
## 3 Other               278
```

```r
Wage %>% split(.$maritl2) %>%
  map(~lm(wage~poly(age,4),data=.))
```

```
## $`1. Never Married`
##
## Call:
## lm(formula = wage ~ poly(age, 4), data = .)
##
## Coefficients:
##   (Intercept)  poly(age, 4)1  poly(age, 4)2  poly(age, 4)3  poly(age, 4)4
##         92.73         217.90        -200.74          97.66         -33.66
##
##
## $`2. Married`
##
## Call:
## lm(formula = wage ~ poly(age, 4), data = .)
##
## Coefficients:
##   (Intercept)  poly(age, 4)1  poly(age, 4)2  poly(age, 4)3  poly(age, 4)4
##        118.86         139.39        -307.10          71.56        -102.80
##
##
## $Other
##
## Call:
## lm(formula = wage ~ poly(age, 4), data = .)
##
```

```r
Wage %>% split(.$maritl2) %>%
  map(~lm(wage~poly(age,4),data=.)) %>%
  map_dbl(~ mean(.$residuals^2))
```

```
## 1. Never Married        2. Married             Other
##          930.132          1796.081          1042.070
```

```r
data(iris)
iris %>% group_by(Species) %>%
  summarize(n=n(),
            meanSL = mean(Sepal.Length),
            meanSW = mean(Sepal.Width),
            SDSL = sd(Sepal.Length),
            SDSW = sd(Sepal.Width))
```

```
## # A tibble: 3 x 6
##   Species        n meanSL meanSW  SDSL  SDSW
##   <fct>      <int>  <dbl>  <dbl> <dbl> <dbl>
## 1 setosa        50   5.01   3.43 0.352 0.379
## 2 versicolor    50   5.94   2.77 0.516 0.314
## 3 virginica     50   6.59   2.97 0.636 0.322
```

```r
set.seed(1)
iris <- iris %>%
  group_by(Species) %>%
  sample_n(size=5) %>%
  ungroup()
library(ggplot2)
ggplot(iris,aes(x=Sepal.Length,y=Sepal.Width,label=Species)) + geom_text()
```