Chapter 10

Miscellaneous examples and data tricks

Reading: Hadley Wickham, "Tidy Data" (Journal of Statistical Software, 2014):

http://vita.had.co.nz/papers/tidy-data.pdf

Warning: This chapter is pretty random and rough.

10.1 Tuberculosis

Data: https://github.com/hadley/tidy-data/blob/master/data/tb.csv

Let's explore tuberculosis in the U.S. How do male and female rates compare?

```
library(plyr)
tb = read.csv("https://github.com/hadley/tidy-data/raw/master/data/tb.csv",
    stringsAsFactors = FALSE)
tb.US = subset(tb, iso2 == "US")
# tb.US = as_tibble(tb.US)
```

Gather up the counts into one column:

```
library(dplyr)
library(tidyr)
tb.long = tb.US %>% gather(demographic, count, -iso2, -year, na.rm = TRUE)
```

Separate sex and age using `separate()`:

```
tb.sep = tb.long %>% separate(demographic, c("sex", "age"), sep = 8)
# Get rid of totals
tb.sep = subset(tb.sep, sex != "new_sp")
tb.sep
```

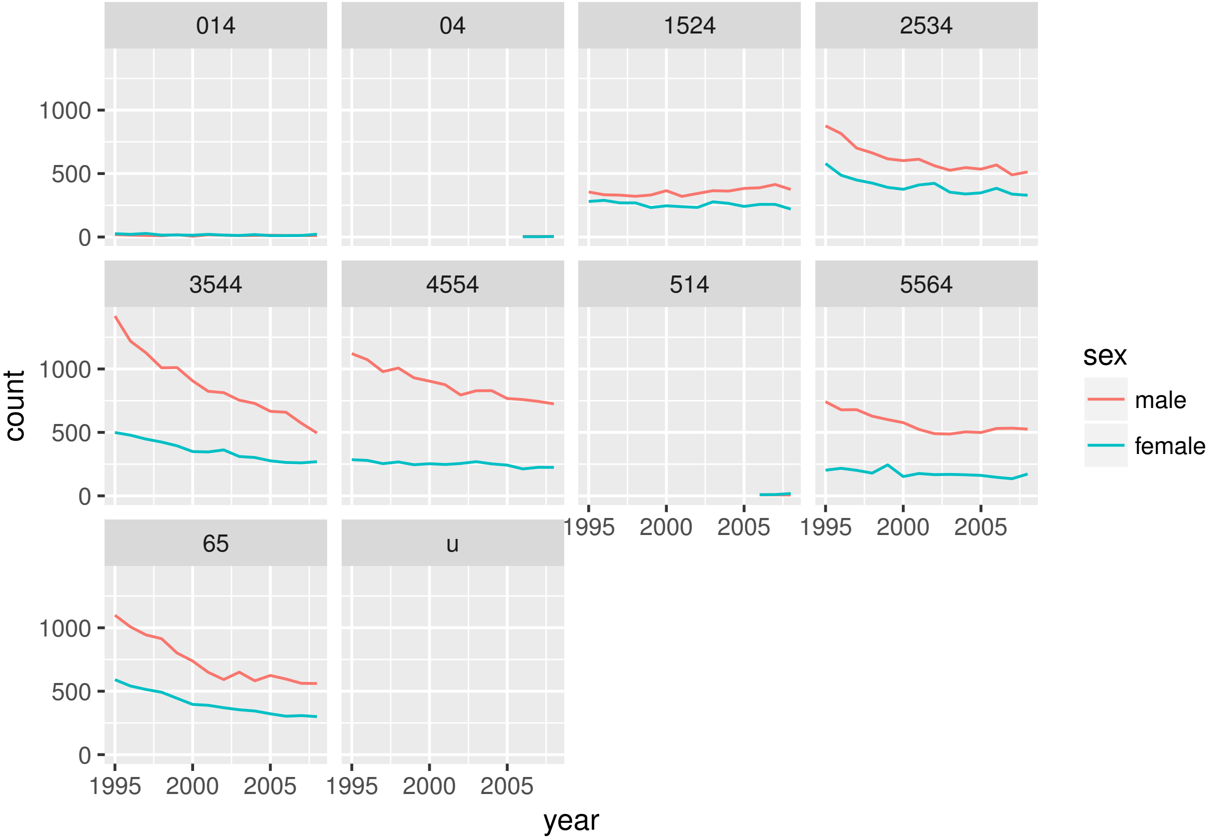
```
##     iso2 year      sex  age count
## 56    US 2006 new_sp_m   04     4
## 57    US 2007 new_sp_m   04     4
## 58    US 2008 new_sp_m   04     4
## 85    US 2006 new_sp_m  514     8
## 86    US 2007 new_sp_m  514     8
## 87    US 2008 new_sp_m  514     7
## 103   US 1995 new_sp_m  014    19
## 104   US 1996 new_sp_m  014    15
## 105   US 1997 new_sp_m  014    12
## 106   US 1998 new_sp_m  014    10
## 107   US 1999 new_sp_m  014    18
## 108   US 2000 new_sp_m  014     6
## 109   US 2001 new_sp_m  014    17
## 110   US 2002 new_sp_m  014    14
## 111   US 2003 new_sp_m  014    11
## 112   US 2004 new_sp_m  014    12
## 113   US 2005 new_sp_m  014    14
## 114   US 2006 new_sp_m  014    12
## 115   US 2007 new_sp_m  014    12
## 116   US 2008 new_sp_m  014    11
## 132   US 1995 new_sp_m 1524   355
## 133   US 1996 new_sp_m 1524   333
## 134   US 1997 new_sp_m 1524   330
## 135   US 1998 new_sp_m 1524   321
## 136   US 1999 new_sp_m 1524   331
## 137   US 2000 new_sp_m 1524   365
## 138   US 2001 new_sp_m 1524   320
## 139   US 2002 new_sp_m 1524   343
## 140   US 2003 new_sp_m 1524   365
## 141   US 2004 new_sp_m 1524   362
## 142   US 2005 new_sp_m 1524   383
## 143   US 2006 new_sp_m 1524   388
## 144   US 2007 new_sp_m 1524   414
## 145   US 2008 new_sp_m 1524   375
## 161   US 1995 new_sp_m 2534   876
## 162   US 1996 new_sp_m 2534   815
## 163   US 1997 new_sp_m 2534   701
## 164   US 1998 new_sp_m 2534   663
## 165   US 1999 new_sp_m 2534   616
## 166   US 2000 new_sp_m 2534   602
## 167   US 2001 new_sp_m 2534   613
## 168   US 2002 new_sp_m 2534   562
## 169   US 2003 new_sp_m 2534   526
## 170   US 2004 new_sp_m 2534   547
## 171   US 2005 new_sp_m 2534   535
## 172   US 2006 new_sp_m 2534   568
## 173   US 2007 new_sp_m 2534   490
## 174   US 2008 new_sp_m 2534   513
## 190   US 1995 new_sp_m 3544  1417
## 191   US 1996 new_sp_m 3544  1219
## 192   US 1997 new_sp_m 3544  1127
## 193   US 1998 new_sp_m 3544  1009
## 194   US 1999 new_sp_m 3544  1011
## 195   US 2000 new_sp_m 3544   906
## 196   US 2001 new_sp_m 3544   824
## 197   US 2002 new_sp_m 3544   813
## 198   US 2003 new_sp_m 3544   754
## 199   US 2004 new_sp_m 3544   728
## 200   US 2005 new_sp_m 3544   666
## 201   US 2006 new_sp_m 3544   659
## 202   US 2007 new_sp_m 3544   572
## 203   US 2008 new_sp_m 3544   495
## 219   US 1995 new_sp_m 4554  1121
## 220   US 1996 new_sp_m 4554  1073
## 221   US 1997 new_sp_m 4554   979
## 222   US 1998 new_sp_m 4554  1007
## 223   US 1999 new_sp_m 4554   930
## 224   US 2000 new_sp_m 4554   904
## 225   US 2001 new_sp_m 4554   876
## 226   US 2002 new_sp_m 4554   795
## 227   US 2003 new_sp_m 4554   828
## 228   US 2004 new_sp_m 4554   829
## 229   US 2005 new_sp_m 4554   767
## 230   US 2006 new_sp_m 4554   759
## 231   US 2007 new_sp_m 4554   744
## 232   US 2008 new_sp_m 4554   725
## 248   US 1995 new_sp_m 5564   742
## 249   US 1996 new_sp_m 5564   678
## 250   US 1997 new_sp_m 5564   679
## 251   US 1998 new_sp_m 5564   628
## 252   US 1999 new_sp_m 5564   601
## 253   US 2000 new_sp_m 5564   577
## 254   US 2001 new_sp_m 5564   524
## 255   US 2002 new_sp_m 5564   490
## 256   US 2003 new_sp_m 5564   487
## 257   US 2004 new_sp_m 5564   504
## 258   US 2005 new_sp_m 5564   499
## 259   US 2006 new_sp_m 5564   531
## 260   US 2007 new_sp_m 5564   533
## 261   US 2008 new_sp_m 5564   526
## 277   US 1995 new_sp_m   65  1099
## 278   US 1996 new_sp_m   65  1007
## 279   US 1997 new_sp_m   65   944
## 280   US 1998 new_sp_m   65   914
## 281   US 1999 new_sp_m   65   801
## 282   US 2000 new_sp_m   65   738
## 283   US 2001 new_sp_m   65   649
## 284   US 2002 new_sp_m   65   592
## 285   US 2003 new_sp_m   65   650
## 286   US 2004 new_sp_m   65   582
## 287   US 2005 new_sp_m   65   624
## 288   US 2006 new_sp_m   65   596
## 289   US 2007 new_sp_m   65   562
## 290   US 2008 new_sp_m   65   561
## 319   US 2008 new_sp_m    u     0
## 346   US 2006 new_sp_f   04     2
## 347   US 2007 new_sp_f   04     2
## 348   US 2008 new_sp_f   04     4
## 375   US 2006 new_sp_f  514     9
## 376   US 2007 new_sp_f  514    10
## 377   US 2008 new_sp_f  514    18
## 393   US 1995 new_sp_f  014    26
## 394   US 1996 new_sp_f  014    21
## 395   US 1997 new_sp_f  014    28
## 396   US 1998 new_sp_f  014    15
## 397   US 1999 new_sp_f  014    16
## 398   US 2000 new_sp_f  014    14
## 399   US 2001 new_sp_f  014    21
## 400   US 2002 new_sp_f  014    15
## 401   US 2003 new_sp_f  014    12
## 402   US 2004 new_sp_f  014    19
## 403   US 2005 new_sp_f  014    11
## 404   US 2006 new_sp_f  014    11
## 405   US 2007 new_sp_f  014    12
## 406   US 2008 new_sp_f  014    22
## 422   US 1995 new_sp_f 1524   280
## 423   US 1996 new_sp_f 1524   289
## 424   US 1997 new_sp_f 1524   269
## 425   US 1998 new_sp_f 1524   269
## 426   US 1999 new_sp_f 1524   232
## 427   US 2000 new_sp_f 1524   246
## 428   US 2001 new_sp_f 1524   239
## 429   US 2002 new_sp_f 1524   233
## 430   US 2003 new_sp_f 1524   277
## 431   US 2004 new_sp_f 1524   265
## 432   US 2005 new_sp_f 1524   241
## 433   US 2006 new_sp_f 1524   257
## 434   US 2007 new_sp_f 1524   257
## 435   US 2008 new_sp_f 1524   220
## 451   US 1995 new_sp_f 2534   579
## 452   US 1996 new_sp_f 2534   487
## 453   US 1997 new_sp_f 2534   449
## 454   US 1998 new_sp_f 2534   425
## 455   US 1999 new_sp_f 2534   391
## 456   US 2000 new_sp_f 2534   376
## 457   US 2001 new_sp_f 2534   410
## 458   US 2002 new_sp_f 2534   423
## 459   US 2003 new_sp_f 2534   353
## 460   US 2004 new_sp_f 2534   339
## 461   US 2005 new_sp_f 2534   348
## 462   US 2006 new_sp_f 2534   384
## 463   US 2007 new_sp_f 2534   338
## 464   US 2008 new_sp_f 2534   329
## 480   US 1995 new_sp_f 3544   499
## 481   US 1996 new_sp_f 3544   478
## 482   US 1997 new_sp_f 3544   447
## 483   US 1998 new_sp_f 3544   424
## 484   US 1999 new_sp_f 3544   394
## 485   US 2000 new_sp_f 3544   349
## 486   US 2001 new_sp_f 3544   346
## 487   US 2002 new_sp_f 3544   362
## 488   US 2003 new_sp_f 3544   310
## 489   US 2004 new_sp_f 3544   302
## 490   US 2005 new_sp_f 3544   276
## 491   US 2006 new_sp_f 3544   263
## 492   US 2007 new_sp_f 3544   260
## 493   US 2008 new_sp_f 3544   269
## 509   US 1995 new_sp_f 4554   285
## 510   US 1996 new_sp_f 4554   279
## 511   US 1997 new_sp_f 4554   254
## 512   US 1998 new_sp_f 4554   267
## 513   US 1999 new_sp_f 4554   245
## 514   US 2000 new_sp_f 4554   253
## 515   US 2001 new_sp_f 4554   247
## 516   US 2002 new_sp_f 4554   255
## 517   US 2003 new_sp_f 4554   269
## 518   US 2004 new_sp_f 4554   252
## 519   US 2005 new_sp_f 4554   242
## 520   US 2006 new_sp_f 4554   212
## 521   US 2007 new_sp_f 4554   225
## 522   US 2008 new_sp_f 4554   224
## 538   US 1995 new_sp_f 5564   202
## 539   US 1996 new_sp_f 5564   217
## 540   US 1997 new_sp_f 5564   201
## 541   US 1998 new_sp_f 5564   179
## 542   US 1999 new_sp_f 5564   244
## 543   US 2000 new_sp_f 5564   152
## 544   US 2001 new_sp_f 5564   176
## 545   US 2002 new_sp_f 5564   167
## 546   US 2003 new_sp_f 5564   169
## 547   US 2004 new_sp_f 5564   166
## 548   US 2005 new_sp_f 5564   161
## 549   US 2006 new_sp_f 5564   146
## 550   US 2007 new_sp_f 5564   135
## 551   US 2008 new_sp_f 5564   172
## 567   US 1995 new_sp_f   65   591
## 568   US 1996 new_sp_f   65   541
## 569   US 1997 new_sp_f   65   514
## 570   US 1998 new_sp_f   65   492
## 571   US 1999 new_sp_f   65   444
## 572   US 2000 new_sp_f   65   396
## 573   US 2001 new_sp_f   65   389
## 574   US 2002 new_sp_f   65   370
## 575   US 2003 new_sp_f   65   354
## 576   US 2004 new_sp_f   65   344
## 577   US 2005 new_sp_f   65   322
## 578   US 2006 new_sp_f   65   303
## 579   US 2007 new_sp_f   65   308
## 580   US 2008 new_sp_f   65   300
## 609   US 2008 new_sp_f    u     0
```

```r
# Rename categories
tb.sep$sex[tb.sep$sex == "new_sp_m"] = "male"
tb.sep$sex[tb.sep$sex == "new_sp_f"] = "female"
tb.sep$sex = factor(tb.sep$sex, levels = c("male", "female"))
tb.sep
```

```
##     iso2 year    sex  age count
## 56    US 2006   male   04     4
## 57    US 2007   male   04     4
## 58    US 2008   male   04     4
## 85    US 2006   male  514     8
## 86    US 2007   male  514     8
## 87    US 2008   male  514     7
## 103   US 1995   male  014    19
## 104   US 1996   male  014    15
## 105   US 1997   male  014    12
## 106   US 1998   male  014    10
## 107   US 1999   male  014    18
## 108   US 2000   male  014     6
## 109   US 2001   male  014    17
## 110   US 2002   male  014    14
## 111   US 2003   male  014    11
## 112   US 2004   male  014    12
## 113   US 2005   male  014    14
## 114   US 2006   male  014    12
## 115   US 2007   male  014    12
## 116   US 2008   male  014    11
## 132   US 1995   male 1524   355
## 133   US 1996   male 1524   333
## 134   US 1997   male 1524   330
## 135   US 1998   male 1524   321
## 136   US 1999   male 1524   331
## 137   US 2000   male 1524   365
## 138   US 2001   male 1524   320
## 139   US 2002   male 1524   343
## 140   US 2003   male 1524   365
## 141   US 2004   male 1524   362
## 142   US 2005   male 1524   383
## 143   US 2006   male 1524   388
## 144   US 2007   male 1524   414
## 145   US 2008   male 1524   375
## 161   US 1995   male 2534   876
## 162   US 1996   male 2534   815
## 163   US 1997   male 2534   701
## 164   US 1998   male 2534   663
## 165   US 1999   male 2534   616
## 166   US 2000   male 2534   602
## 167   US 2001   male 2534   613
## 168   US 2002   male 2534   562
## 169   US 2003   male 2534   526
## 170   US 2004   male 2534   547
## 171   US 2005   male 2534   535
## 172   US 2006   male 2534   568
## 173   US 2007   male 2534   490
## 174   US 2008   male 2534   513
## 190   US 1995   male 3544  1417
## 191   US 1996   male 3544  1219
## 192   US 1997   male 3544  1127
## 193   US 1998   male 3544  1009
## 194   US 1999   male 3544  1011
## 195   US 2000   male 3544   906
## 196   US 2001   male 3544   824
## 197   US 2002   male 3544   813
## 198   US 2003   male 3544   754
## 199   US 2004   male 3544   728
## 200   US 2005   male 3544   666
## 201   US 2006   male 3544   659
## 202   US 2007   male 3544   572
## 203   US 2008   male 3544   495
## 219   US 1995   male 4554  1121
## 220   US 1996   male 4554  1073
## 221   US 1997   male 4554   979
## 222   US 1998   male 4554  1007
## 223   US 1999   male 4554   930
## 224   US 2000   male 4554   904
## 225   US 2001   male 4554   876
## 226   US 2002   male 4554   795
## 227   US 2003   male 4554   828
## 228   US 2004   male 4554   829
## 229   US 2005   male 4554   767
## 230   US 2006   male 4554   759
## 231   US 2007   male 4554   744
## 232   US 2008   male 4554   725
## 248   US 1995   male 5564   742
## 249   US 1996   male 5564   678
## 250   US 1997   male 5564   679
## 251   US 1998   male 5564   628
## 252   US 1999   male 5564   601
## 253   US 2000   male 5564   577
## 254   US 2001   male 5564   524
## 255   US 2002   male 5564   490
## 256   US 2003   male 5564   487
## 257   US 2004   male 5564   504
## 258   US 2005   male 5564   499
## 259   US 2006   male 5564   531
## 260   US 2007   male 5564   533
## 261   US 2008   male 5564   526
## 277   US 1995   male   65  1099
## 278   US 1996   male   65  1007
## 279   US 1997   male   65   944
## 280   US 1998   male   65   914
## 281   US 1999   male   65   801
## 282   US 2000   male   65   738
## 283   US 2001   male   65   649
## 284   US 2002   male   65   592
## 285   US 2003   male   65   650
## 286   US 2004   male   65   582
## 287   US 2005   male   65   624
## 288   US 2006   male   65   596
## 289   US 2007   male   65   562
## 290   US 2008   male   65   561
## 319   US 2008   male    u     0
## 346   US 2006 female   04     2
## 347   US 2007 female   04     2
## 348   US 2008 female   04     4
## 375   US 2006 female  514     9
## 376   US 2007 female  514    10
## 377   US 2008 female  514    18
## 393   US 1995 female  014    26
## 394   US 1996 female  014    21
## 395   US 1997 female  014    28
## 396   US 1998 female  014    15
## 397   US 1999 female  014    16
## 398   US 2000 female  014    14
## 399   US 2001 female  014    21
## 400   US 2002 female  014    15
## 401   US 2003 female  014    12
## 402   US 2004 female  014    19
## 403   US 2005 female  014    11
## 404   US 2006 female  014    11
## 405   US 2007 female  014    12
## 406   US 2008 female  014    22
## 422   US 1995 female 1524   280
## 423   US 1996 female 1524   289
## 424   US 1997 female 1524   269
## 425   US 1998 female 1524   269
## 426   US 1999 female 1524   232
## 427   US 2000 female 1524   246
## 428   US 2001 female 1524   239
## 429   US 2002 female 1524   233
## 430   US 2003 female 1524   277
## 431   US 2004 female 1524   265
## 432   US 2005 female 1524   241
## 433   US 2006 female 1524   257
## 434   US 2007 female 1524   257
## 435   US 2008 female 1524   220
## 451   US 1995 female 2534   579
## 452   US 1996 female 2534   487
## 453   US 1997 female 2534   449
## 454   US 1998 female 2534   425
## 455   US 1999 female 2534   391
## 456   US 2000 female 2534   376
## 457   US 2001 female 2534   410
## 458   US 2002 female 2534   423
## 459   US 2003 female 2534   353
## 460   US 2004 female 2534   339
## 461   US 2005 female 2534   348
## 462   US 2006 female 2534   384
## 463   US 2007 female 2534   338
## 464   US 2008 female 2534   329
## 480   US 1995 female 3544   499
## 481   US 1996 female 3544   478
## 482   US 1997 female 3544   447
## 483   US 1998 female 3544   424
## 484   US 1999 female 3544   394
## 485   US 2000 female 3544   349
## 486   US 2001 female 3544   346
## 487   US 2002 female 3544   362
## 488   US 2003 female 3544   310
## 489   US 2004 female 3544   302
## 490   US 2005 female 3544   276
## 491   US 2006 female 3544   263
## 492   US 2007 female 3544   260
## 493   US 2008 female 3544   269
## 509   US 1995 female 4554   285
## 510   US 1996 female 4554   279
## 511   US 1997 female 4554   254
## 512   US 1998 female 4554   267
## 513   US 1999 female 4554   245
## 514   US 2000 female 4554   253
## 515   US 2001 female 4554   247
## 516   US 2002 female 4554   255
## 517   US 2003 female 4554   269
## 518   US 2004 female 4554   252
## 519   US 2005 female 4554   242
## 520   US 2006 female 4554   212
## 521   US 2007 female 4554   225
## 522   US 2008 female 4554   224
## 538   US 1995 female 5564   202
## 539   US 1996 female 5564   217
## 540   US 1997 female 5564   201
## 541   US 1998 female 5564   179
## 542   US 1999 female 5564   244
## 543   US 2000 female 5564   152
## 544   US 2001 female 5564   176
## 545   US 2002 female 5564   167
## 546   US 2003 female 5564   169
## 547   US 2004 female 5564   166
## 548   US 2005 female 5564   161
## 549   US 2006 female 5564   146
## 550   US 2007 female 5564   135
## 551   US 2008 female 5564   172
## 567   US 1995 female   65   591
## 568   US 1996 female   65   541
## 569   US 1997 female   65   514
## 570   US 1998 female   65   492
## 571   US 1999 female   65   444
## 572   US 2000 female   65   396
## 573   US 2001 female   65   389
## 574   US 2002 female   65   370
## 575   US 2003 female   65   354
## 576   US 2004 female   65   344
## 577   US 2005 female   65   322
## 578   US 2006 female   65   303
## 579   US 2007 female   65   308
## 580   US 2008 female   65   300
## 609   US 2008 female    u     0
```

Answer the question:

```
library(ggplot2)
ggplot(tb.sep, aes(x = year, y = count, color = sex)) + geom_line() + facet_wrap(~age)
```



In all age groups with non-negligible counts, men have higher TB rates than women.

10.2 Baseball

```
# install.packages('Lahman')
library(Lahman)
```

Batting data:

```
# Batting = as_tibble(Batting)
dim(Batting)
```

```
## [1] 101332     22
```

```
names(Batting)
```

```
##  [1] "playerID" "yearID"   "stint"    "teamID"   "lgID"     "G"
##  [7] "AB"       "R"        "H"        "X2B"      "X3B"      "HR"
## [13] "RBI"      "SB"       "CS"       "BB"       "SO"       "IBB"
## [19] "HBP"      "SH"       "SF"       "GIDP"
```

Our goal is to predict a player's 2015 statistics based on 2012 to 2014 statistics. Get 2012–2015 data:

```
pred.year = 2015
B = subset(Batting, yearID >= pred.year - 3 & yearID <= pred.year)
```

Add a variable for "plate appearances":

```
B = transform(B, PA = AB + BB + HBP + SF + SH)
```

Aggregate by year. We'll use `ddply()` from the `plyr` library (there are also functions to do this in `dplyr` but I haven't learned them.)

```
stats = c("PA", "AB", "R", "H", "X2B", "X3B", "HR", "RBI", "SB", "CS", "BB",
    "SO", "IBB", "HBP", "SH", "SF", "GIDP")
B = ddply(B[, c("playerID", "yearID", stats)], ~playerID + yearID, summarise,
    PA = sum(PA), AB = sum(AB), R = sum(R), H = sum(H), X2B = sum(X2B), X3B = sum(X3B),
    HR = sum(HR), RBI = sum(RBI), SB = sum(SB), CS = sum(CS), BB = sum(BB),
    SO = sum(SO), IBB = sum(IBB), HBP = sum(HBP), SH = sum(SH), SF = sum(SF),
    GIDP = sum(GIDP))
dim(B)
```

```
## [1] 5256   19
```

```
summary(B)
```

```
##    playerID             yearID           PA               AB     
##  Length:5256        Min.   :2012   Min.   :  0.0   Min.   :  0  
##  Class :character   1st Qu.:2013   1st Qu.:  0.0   1st Qu.:  0  
##  Mode  :character   Median :2014   Median : 20.0   Median : 18  
##                     Mean   :2014   Mean   :140.1   Mean   :126  
##                     3rd Qu.:2015   3rd Qu.:225.0   3rd Qu.:204  
##                     Max.   :2015   Max.   :740.0   Max.   :684  
##        R                H               X2B              X3B        
##  Min.   :  0.00   Min.   :  0.00   Min.   : 0.000   Min.   : 0.0000  
##  1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.: 0.000   1st Qu.: 0.0000  
##  Median :  1.00   Median :  3.00   Median : 0.000   Median : 0.0000  
##  Mean   : 15.54   Mean   : 31.94   Mean   : 6.252   Mean   : 0.6634  
##  3rd Qu.: 22.00   3rd Qu.: 48.00   3rd Qu.: 9.000   3rd Qu.: 0.0000  
##  Max.   :129.00   Max.   :225.00   Max.   :55.000   Max.   :15.0000  
##        HR              RBI               SB               CS         
##  Min.   : 0.000   Min.   :  0.00   Min.   : 0.000   Min.   : 0.0000  
##  1st Qu.: 0.000   1st Qu.:  0.00   1st Qu.: 0.000   1st Qu.: 0.0000  
##  Median : 0.000   Median :  1.00   Median : 0.000   Median : 0.0000  
##  Mean   : 3.556   Mean   : 14.78   Mean   : 2.129   Mean   : 0.8071  
##  3rd Qu.: 4.000   3rd Qu.: 20.00   3rd Qu.: 1.000   3rd Qu.: 1.0000  
##  Max.   :53.000   Max.   :139.00   Max.   :64.000   Max.   :23.0000  
##        BB               SO               IBB               HBP        
##  Min.   :  0.00   Min.   :  0.00   Min.   : 0.0000   Min.   : 0.000  
##  1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.: 0.0000   1st Qu.: 0.000  
##  Median :  1.00   Median :  6.00   Median : 0.0000   Median : 0.000  
##  Mean   : 10.93   Mean   : 28.16   Mean   : 0.7627   Mean   : 1.196  
##  3rd Qu.: 15.00   3rd Qu.: 45.00   3rd Qu.: 0.0000   3rd Qu.: 1.000  
##  Max.   :143.00   Max.   :222.00   Max.   :29.0000   Max.   :30.000  
##        SH               SF              GIDP       
##  Min.   : 0.000   Min.   : 0.000   Min.   : 0.000  
##  1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.000  
##  Median : 0.000   Median : 0.000   Median : 0.000  
##  Mean   : 1.028   Mean   : 0.942   Mean   : 2.796  
##  3rd Qu.: 1.000   3rd Qu.: 1.000   3rd Qu.: 4.000  
##  Max.   :17.000   Max.   :12.000   Max.   :31.000  
```

Pick out years:

```
B.short = B[, c(1, 2, 3, 4, 5, 6, 9, 10, 13, 14)]
names(B.short)
```

```
##  [1] "playerID" "yearID"   "PA"       "AB"       "R"        "H"
##  [7] "HR"       "RBI"      "BB"       "SO"
```

```
B.2012 = B.short[B.short$yearID == 2012, ]
B.2013 = B.short[B.short$yearID == 2013, ]
B.2014 = B.short[B.short$yearID == 2014, ]
B.2015 = B.short[B.short$yearID == 2015, ]
```

Add names to things:

```
names(B.2012) = c("playerID", "yearID.1", "PA.1", "AB.1", "R.1", "H.1", "HR.1",
    "RBI.1", "BB.1", "SO.1")
names(B.2013) = c("playerID", "yearID.2", "PA.2", "AB.2", "R.2", "H.2", "HR.2",
    "RBI.2", "BB.2", "SO.2")
names(B.2014) = c("playerID", "yearID.3", "PA.3", "AB.3", "R.3", "H.3", "HR.3",
    "RBI.3", "BB.3", "SO.3")
names(B.2015) = c("playerID", "yearID.4", "PA.4", "AB.4", "R.4", "H.4", "HR.4",
    "RBI.4", "BB.4", "SO.4")
```

10.2.1 Merging things

For easy problems `cbind()` and `rbind()` may suffice, but this is harder. Instead, we'll `merge()` everything:

```
B2 = merge(B.2012, B.2013, by = "playerID")
B3 = merge(B2, B.2014, by = "playerID")
B4 = merge(B3, B.2015, by = "playerID")
```

Only keep the complete cases:

```
B.complete = B4[complete.cases(B4), ]
PAs = (B.complete$PA.1 > 0) * (B.complete$PA.2 > 0) * (B.complete$PA.3 > 0) *
    (B.complete$PA.4 > 0)
B.all = B.complete[PAs == 1, ]
```

What if we wanted to include birth year? (We don't need it here, but age would be the next variable that I'd try to include in a model.) Let's add that in from the `Master` data frame:

```
birth = Master[, c("playerID", "birthYear")]
B.merged = merge(B.all, birth, by = "playerID")
```

10.2.2 Making a model

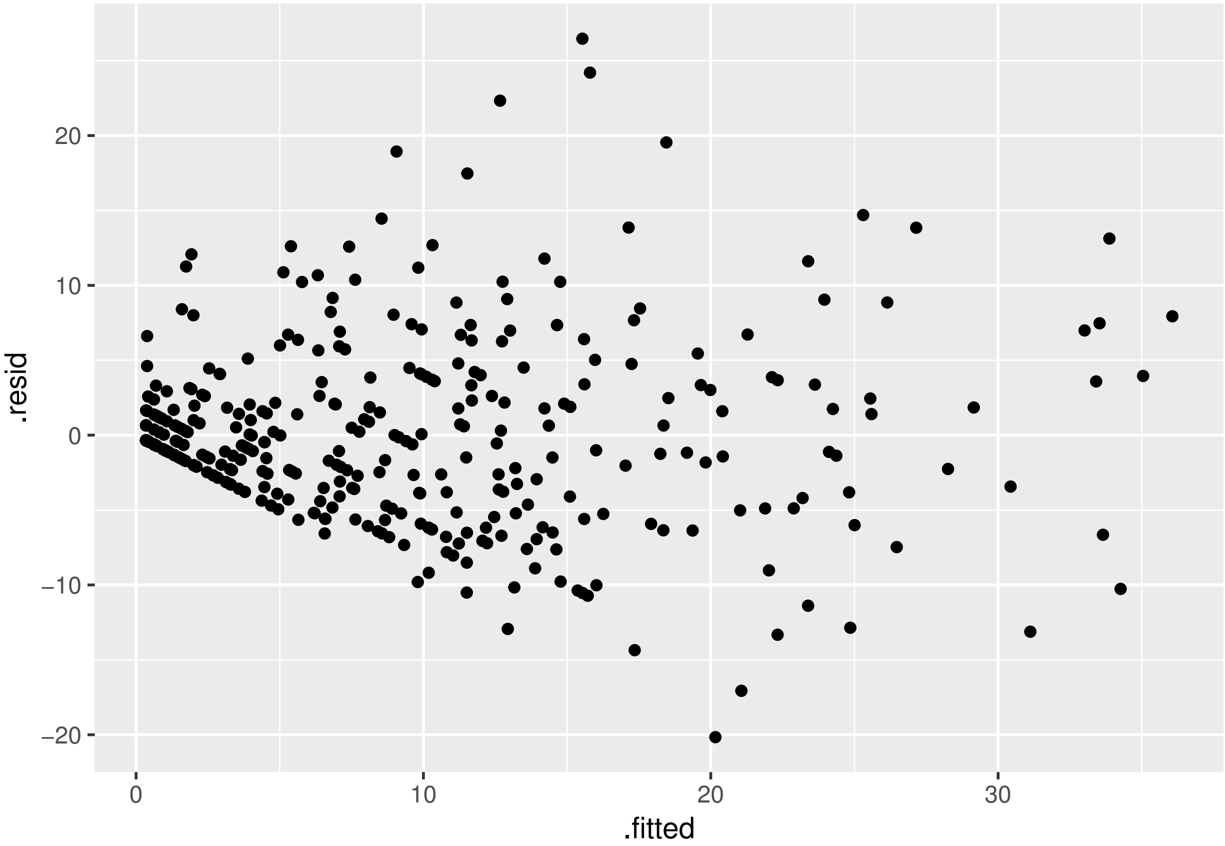
Let's predict home runs based on their past three years of home runs:

```
homerun.lm = lm(HR.4 ~ HR.1 + HR.2 + HR.3, data = B.merged)
library(broom)
tidy(homerun.lm)
```

```
##          term  estimate  std.error  statistic      p.value
## 1 (Intercept) 0.3487871 0.36773423  0.9484760 3.434078e-01
## 2        HR.1 0.0358215 0.04276830  0.8375714 4.027262e-01
## 3        HR.2 0.2732422 0.05269861  5.1849986 3.300467e-07
## 4        HR.3 0.6869728 0.04902130 14.0137630 3.751616e-37
```

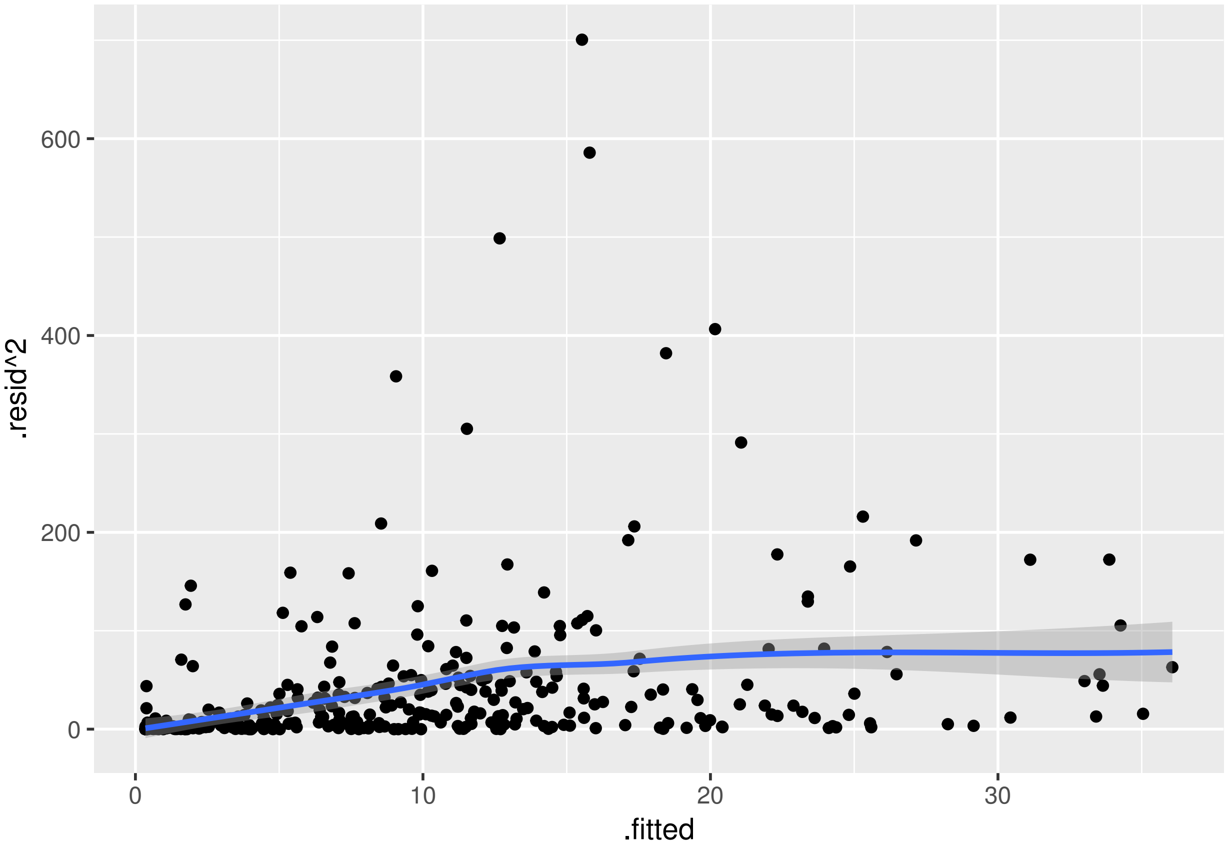
The coefficients get smaller for older years, which makes sense. Look at the residuals:

```
homerun.df = augment(homerun.lm)
ggplot(homerun.df, aes(x = .fitted, y = .resid)) + geom_point()
```



It's heteroskedastic. How do the squared residuals change?

```
homerun.df = augment(homerun.lm)
ggplot(homerun.df, aes(x = .fitted, y = .resid^2)) + geom_point() + geom_smooth(method.args = list(degr
```

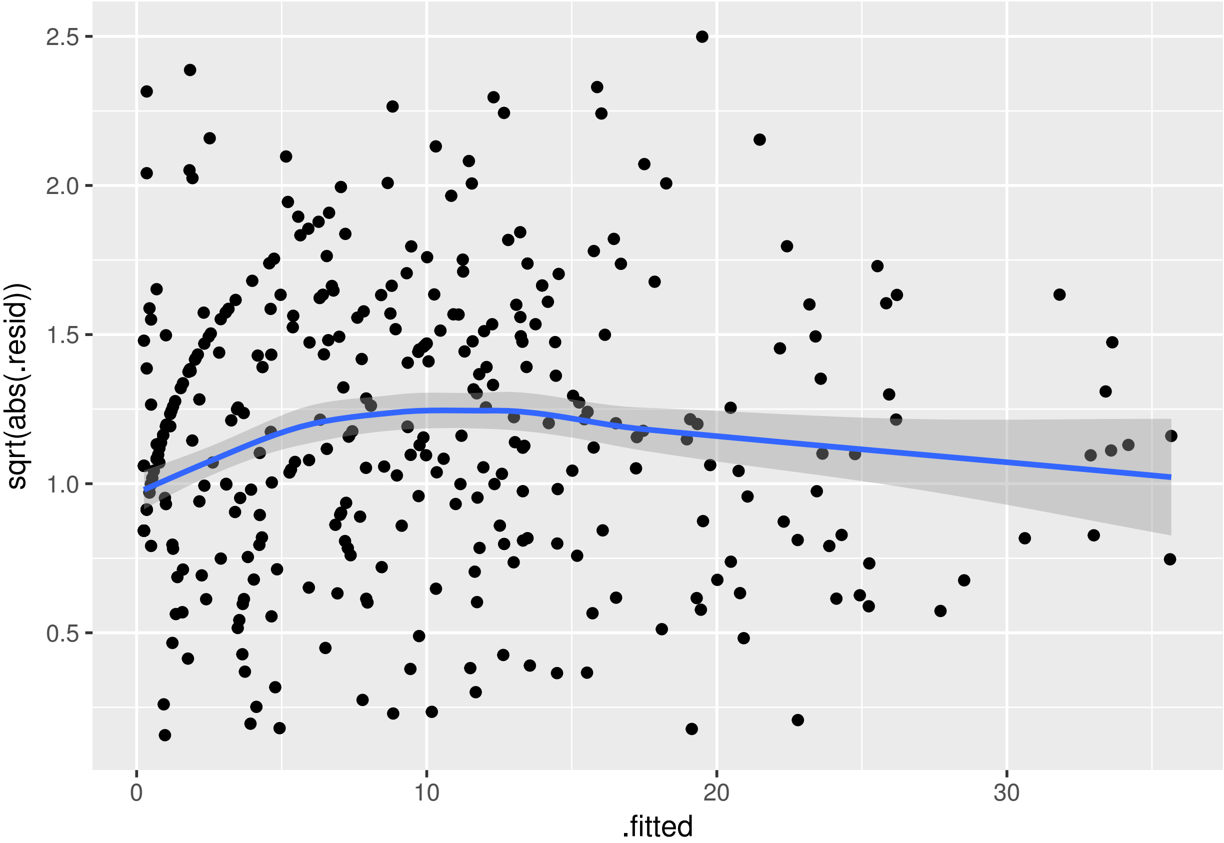


An overdispersed Poisson might work okay. It doesn't make much sense to predict on the log scale, so we fit on the original scale instead.

```
homerun.quasi = glm(HR.4 ~ HR.1 + HR.2 + HR.3, start = coef(homerun.lm), family = quasipoisson(link = "
    data = B.merged)
tidy(homerun.quasi)
```

```
##          term   estimate  std.error statistic      p.value
## 1 (Intercept) 0.25183823 0.09422345  2.672777 7.802190e-03
## 2        HR.1 0.09569843 0.04704579  2.034155 4.253523e-02
## 3        HR.2 0.24533197 0.05891869  4.163907 3.765670e-05
## 4        HR.3 0.66236049 0.06091515 10.873493 1.491511e-24
```

```
homerun.quasi.df = augment(homerun.quasi)
ggplot(homerun.quasi.df, aes(x = .fitted, y = sqrt(abs(.resid)))) + geom_point() +
    geom_smooth(method.args = list(degree = 1))
```



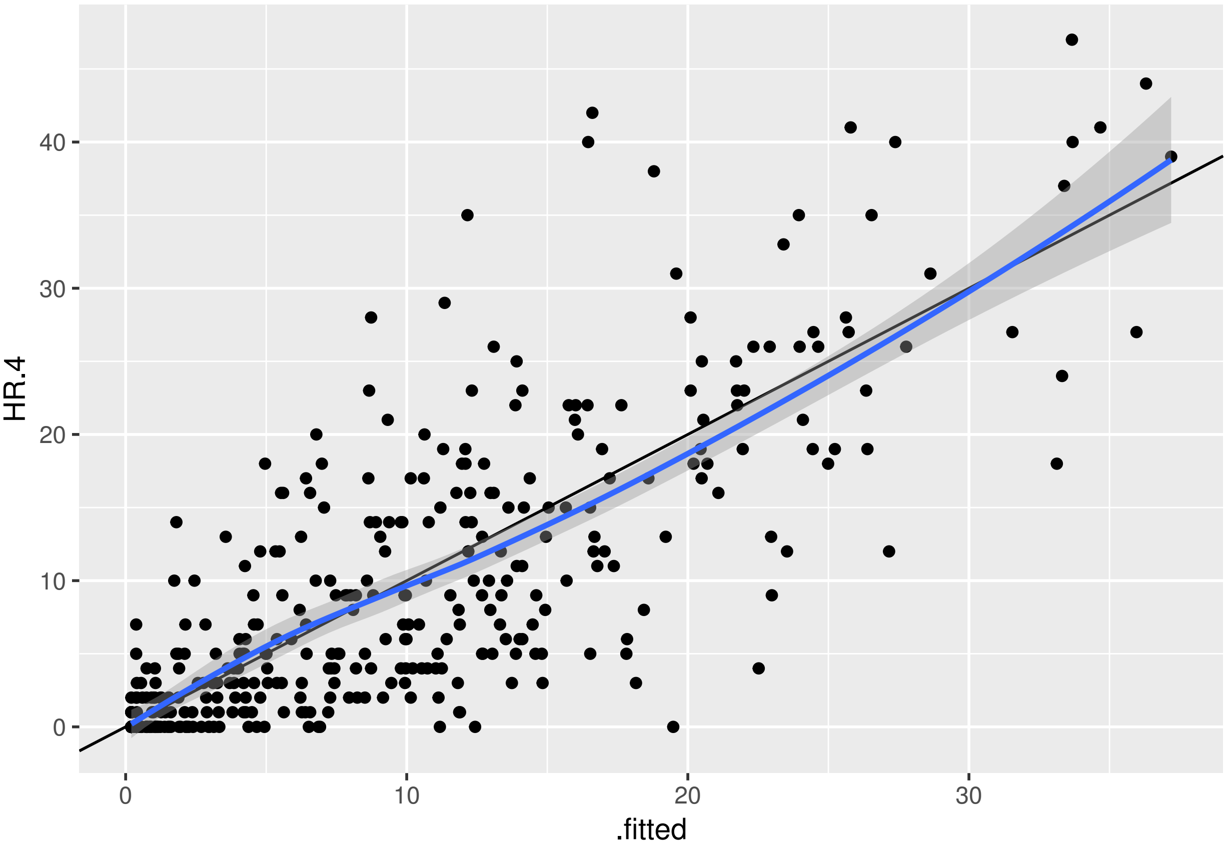
The coefficients are a bit different and the standard errors are a bit bigger. The deviance residuals aren't perfect but they could be worse.

Alternatively, try a negative binomial, which can give probabilistic predictions:

```
library(MASS)
homerun.nb = glm.nb(HR.4 ~ HR.1 + HR.2 + HR.3, data = B.merged, start = coef(homerun.lm),
    link = identity)
tidy(homerun.nb)
```

```
##          term  estimate  std.error statistic      p.value
## 1 (Intercept) 0.2018393 0.04744128  4.254507 2.095097e-05
## 2        HR.1 0.1729842 0.04796215  3.606681 3.101384e-04
## 3        HR.2 0.1971358 0.05634843  3.498515 4.678576e-04
## 4        HR.3 0.6656151 0.06987018  9.526455 1.627453e-21
```

```
homerun.nb.df = augment(homerun.nb, type.residuals = "response")
ggplot(homerun.nb.df, aes(x = .fitted, y = HR.4)) + geom_point() + geom_abline() +
    geom_smooth()
```



10.3 Catterplots

These are self-explanatory I think.

```r
# library(devtools) install_github('Gibbsdavidl/CatterPlots')
library(CatterPlots)
cat.random = multicat(xs = runif(42), ys = runif(42), cat = 1:11, catcolor = list(c(0,
    0, 0, 1)), main = "42 random cats")
```