Homework 1 Solutions

Contents

Load some handy packages:

```
suppressPackageStartupMessages(library(dplyr))
suppressPackageStartupMessages(library(ggplot2))
```

Births

```
#load the fivethirtyeight library
suppressPackageStartupMessages(library(fivethirtyeight))

#load the births data
data(US_births_2000_2014)
```

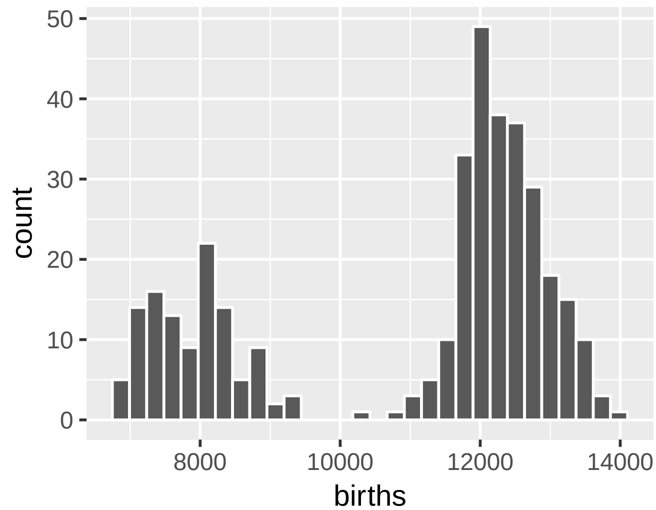
Exercise 1

a. unit of observation = 1 day
b. .

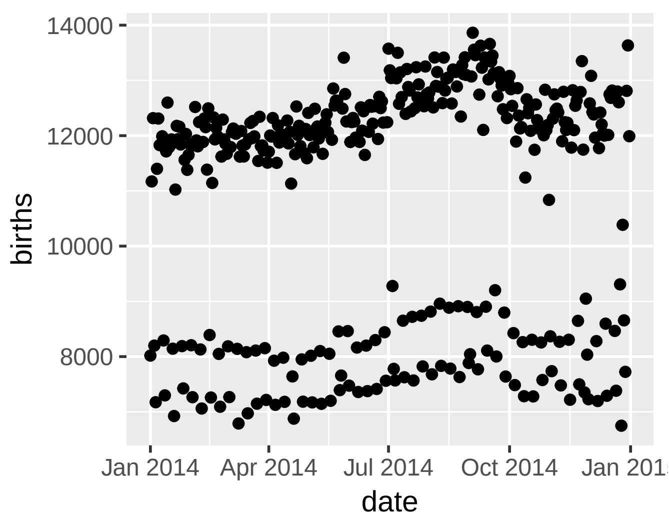
```
dim(US_births_2000_2014)
## [1] 5479    6
```

Exercise 2

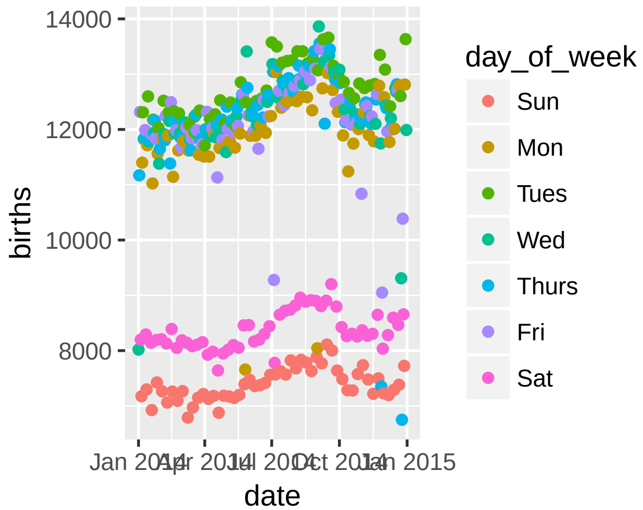
```r
#a.
Births2014 <- US_births_2000_2014 %>%
    filter(year == 2014)

#b.
ggplot(Births2014, aes(x=births)) +
    geom_histogram(color="white")
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```r
#c.
ggplot(Births2014, aes(x=date, y=births)) +
    geom_point()
```




```r
#d.
ggplot(Births2014, aes(x=date, y=births, color=day_of_week)) +
    geom_point()
```



e:

These are holidays (eg: July 4, Thanksgiving, etc)

f:

Births increase in early Fall. Births are less likely to occur on weekends & holidays.

Exercise 3

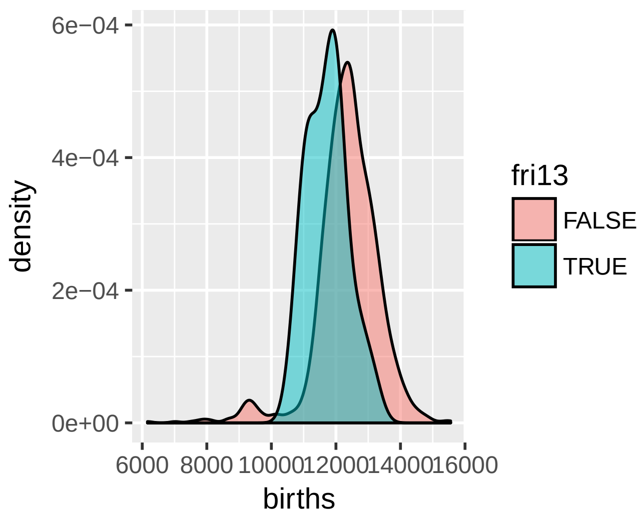
```
allyears <- full_join(US_births_1994_2003, US_births_2000_2014)
## Joining, by = c("year", "month", "date_of_month", "date", "day_of_week", "births")

#a
ggplot(allyears, aes(x=date, y=births, color=day_of_week)) +
     geom_point()
```


   b.

The discrepancy in weekday/weekend births increased over time. The number of births seemed to peak around 2008.

Exercise 4

```
#a
frionly <- allyears %>%                                  #start with all of the data
    filter(day_of_week=="Fri") %>%                       #filter out just the Fridays
    mutate(fri13 = (date_of_month == 13))

#b
#There tend to be fewer births on Friday the 13th than on other Fridays.
ggplot(frionly, aes(x=births, fill=fri13)) +
    geom_density(alpha=0.5)
```

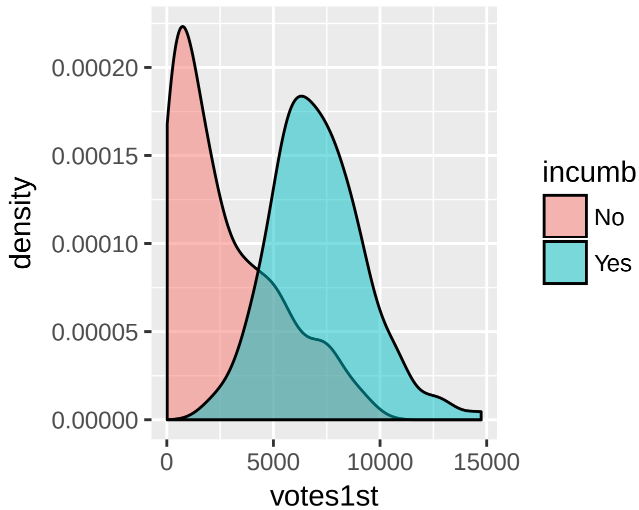


Interaction

```
campaigns = read.csv("https://www.macalester.edu/~ajohns24/data/CampaignSpending.csv")
```

Exercise 5

```
#a
ggplot(campaigns, aes(x=votes1st, fill=incumb)) +
    geom_density(alpha=0.5)
```




```
#b
campaigns %>%
    group_by(incumb) %>%
    summarize(means=mean(votes1st, na.rm=TRUE))
## # A tibble: 2 x 2
##   incumb means
##   <fctr> <dbl>
## 1     No  2722
## 2    Yes  7083

#c
model1 = lm(votes1st ~ incumb, campaigns)
summary(model1)
##
## Call:
## lm(formula = votes1st ~ incumb, data = campaigns)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5075  -1821   -599   1440   7659
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2722        132    20.7   <2e-16 ***
## incumbYes       4361        242    18.0   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2380 on 461 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.414,  Adjusted R-squared:  0.412
## F-statistic:  325 on 1 and 461 DF,  p-value: <2e-16
```

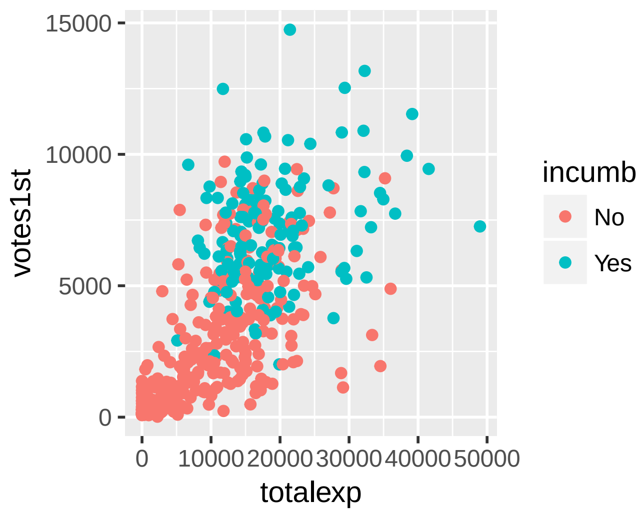
Part c:

votes1st = 2722.0123 + 4361.2286 incumb

On average, challengers receive 2722.0123 votes and incumbents earn 4361.2286 more.

Exercise 6

```r
#a
ggplot(campaigns, aes(x=totalexp, y=votes1st, color=incumb)) +
    geom_point()
```




```r
#b
#full model: votes1st = 1031 + 0.1745 totalexp + 2764 incumbYes
#challenger model: votes1st = 1031 + 0.1745 totalexp
#incumbent model: votes1st = 3795 + 0.1745 totalexp
model2 <- lm(votes1st ~ totalexp + incumb, campaigns)
summary(model2)
##
## Call:
## lm(formula = votes1st ~ totalexp + incumb, data = campaigns)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5259  -1132   -433   1043   7207
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.03e+03   1.57e+02    6.55  1.5e-10 ***
## totalexp    1.74e-01   1.18e-02   14.80  < 2e-16 ***
## incumbYes   2.76e+03   2.27e+02   12.20  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1960 on 460 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.603,  Adjusted R-squared:  0.601
## F-statistic:  349 on 2 and 460 DF,  p-value: <2e-16

#c
#see below

#d
```

```r
1031 + 0.1745*10000 + 2764*0
## [1] 2776
1031 + 0.1745*10000 + 2764*1
## [1] 5540

#e
suppressPackageStartupMessages(library(mosaic))
model2_pred <- makeFun(model2)
model2_pred(incumb="No", totalexp=10000)
##    1
## 2776
model2_pred(incumb="Yes", totalexp=10000)
##    1
## 5540
```

c:

The intercept isn't very meaningful: this would be the predicted number of votes for a theoretical challenger that spends $0.

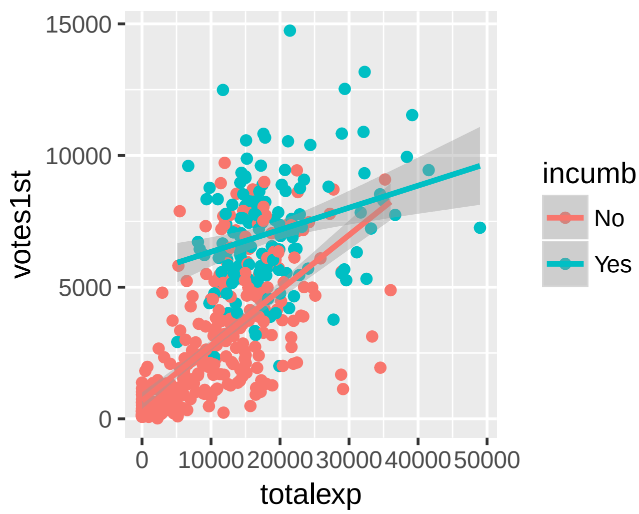
totalexp coef: Controlling for incumbency status, every extra 1,000 euros spent corresponds to a 174.5 vote increase (on average).

incumbYes coef: Controlling for total expenditures, incumbents get 2764 more votes on average than challengers

Exercise 7

```r
newmodel <- lm(votes1st ~ totalexp * incumb, campaigns)
summary(newmodel)
##
## Call:
## lm(formula = votes1st ~ totalexp * incumb, data = campaigns)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5990  -1059   -329    918   7442
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          690.5169   168.6019    4.10  5.0e-05 ***
## totalexp               0.2097     0.0135   15.47  < 2e-16 ***
## incumbYes           4813.8932   472.4071   10.19  < 2e-16 ***
## totalexp:incumbYes   -0.1259     0.0256   -4.91  1.3e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1910 on 459 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.623,  Adjusted R-squared:  0.62
## F-statistic:  252 on 3 and 459 DF,  p-value: <2e-16

#a.
#full: votes1st = 690.5169 + 0.2097 totalexp + 4813.8932 incumbYes - 0.1259 totalexp * incumbYes
```

```r
#challengers: votes1st = 690.5169 + 0.2097 totalexp
#incumbents: votes1st = 5504.41 + 0.0838 totalexp

#b.
690.5169 + 0.2097*10000 + 4813.8932*0 - 0.1259*10000*0
## [1] 2788
690.5169 + 0.2097*10000 + 4813.8932*1 - 0.1259*10000*1
## [1] 6342
newmodel_pred <- makeFun(newmodel)
newmodel_pred(incumb="No", totalexp=10000)
##    1
## 2787
newmodel_pred(incumb="Yes", totalexp=10000)
##    1
## 6342

#c
#Challengers enjoy a greater return on spending than challengers do
ggplot(campaigns, aes(x=totalexp, y=votes1st, col=incumb)) +
    geom_point() +
    geom_smooth(method="lm")
```




```r
#d
#see below
```

d:

- Intercept: This is the intercept for challengers. On average, challengers that spend 0 euros receive 691 votes.

- totalexp: This is the slope for challengers. On average, every extra 100 euros spent corresponds to a 210 vote increase for challengers.
- incumbYes: This is the change in intercept for incumbents. On average, incumbents that spend 0 euros will receive 4814 more votes than a challenger that spends 0 euros.
- interaction: This is the change in slope for incumbents. On average, the increase in votes corresponding to a 100 euros increase in spending is 126 votes less for incumbents than for challengers.

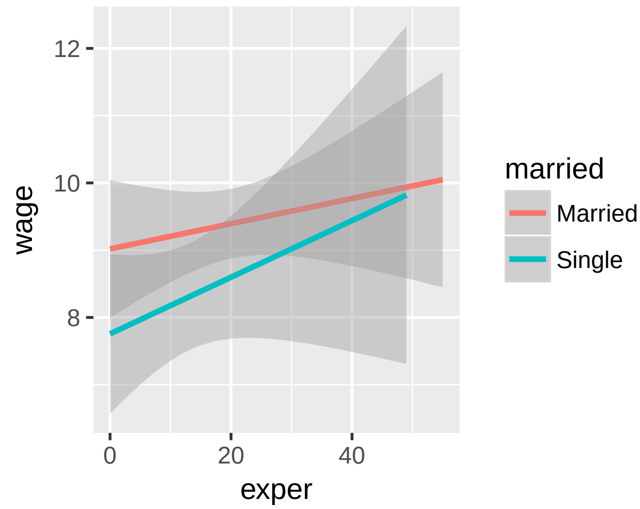
Covariates

```
#Load the data:
suppressPackageStartupMessages(library(mosaic))
data(CPS85)
head(CPS85,3)
##   wage educ race sex hispanic south married exper union age sector
## 1  9.0   10    W   M       NH    NS Married    27   Not  43  const
## 2  5.5   12    W   M       NH    NS Married    20   Not  38  sales
## 3  3.8   12    W   F       NH    NS  Single     4   Not  22  sales
```

Exercise 8

Exercise 9

```
#a
cpsmod2 <- lm(wage ~ married*exper, data=CPS85)
summary(cpsmod2)
##
## Call:
## lm(formula = wage ~ married * exper, data = CPS85)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8.47  -3.75  -1.24   2.15  36.70
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           9.0196     0.5399   16.71   <2e-16 ***
## marriedSingle        -1.2634     0.7798   -1.62     0.11
## exper                 0.0187     0.0230    0.81     0.42
## marriedSingle:exper   0.0234     0.0391    0.60     0.55
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.12 on 530 degrees of freedom
## Multiple R-squared:  0.0146, Adjusted R-squared:  0.00906
## F-statistic: 2.63 on 3 and 530 DF,  p-value: 0.0498

#b
#The increase in wage with experience is more rapid for single vs married workers
ggplot(CPS85, aes(y=wage, x=exper, color=married)) +
    geom_smooth(method="lm")
```




```r
#c
#married
9.01957 - 1.26343*0 + 0.01871*10 + 0.02339*0*10
## [1] 9.207
#single
9.01957 - 1.26343*1 + 0.01871*10 + 0.02339*1*10
## [1] 8.177
#difference
9.20667 - 8.17714
## [1] 1.03

#d
#married
9.01957 - 1.26343*0 + 0.01871*20 + 0.02339*0*20
## [1] 9.394
#single
9.01957 - 1.26343*1 + 0.01871*20 + 0.02339*1*20
## [1] 8.598
#difference
9.39377 - 8.59814
## [1] 0.7956
```

Exercise 10

```r
cpsmod3 <- lm(wage ~ exper + educ + sector + married, data=CPS85)
summary(cpsmod3)
##
## Call:
## lm(formula = wage ~ exper + educ + sector + married, data = CPS85)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -10.46  -2.84  -0.70   1.87  34.12
##
## Coefficients:
```

```
## (Intercept)    -3.7517     1.5348   -2.44   0.0148 *
## exper           0.0944     0.0175    5.40  1.0e-07 ***
## educ            0.7474     0.1010    7.40  5.4e-13 ***
## sectorconst     3.0042     1.0977    2.74   0.0064 **
## sectormanag     3.9732     0.7651    5.19  3.0e-07 ***
## sectormanuf     1.6723     0.7188    2.33   0.0204 *
## sectorother     2.1319     0.7109    3.00   0.0028 **
## sectorprof      2.6686     0.6763    3.95  9.0e-05 ***
## sectorsales    -0.1774     0.8482   -0.21   0.8344
## sectorservice  -0.1218     0.6725   -0.18   0.8564
## marriedSingle  -0.3985     0.4221   -0.94   0.3456
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.42 on 523 degrees of freedom
## Multiple R-squared:  0.275,  Adjusted R-squared:  0.261
## F-statistic: 19.9 on 10 and 523 DF,  p-value: <2e-16

#a
#16 parallel planes

#b
#married
-3.75175 + 0.09444*10 + 0.74744*16 - 0.12179*1 - 0.39846*0
## [1] 9.03
#single
-3.75175 + 0.09444*10 + 0.74744*16 - 0.12179*1 - 0.39846*1
## [1] 8.631
#difference
9.0299 - 8.63144
## [1] 0.3985

#c
#married
-3.75175 + 0.09444*20 + 0.74744*12 + 1.67234*1 - 0.39846*0
## [1] 8.779
#single
-3.75175 + 0.09444*20 + 0.74744*12 + 1.67234*1 - 0.39846*1
## [1] 8.38
#difference
8.77867 - 8.38021
## [1] 0.3985

#d & e
# see below
```

d. When controlling / holding constant experience, education, and job sector, single people make 40 cents less per hour than married people (on average).

e. In the first model we weren't controlling for any labor covariates.

Exercise 11

```
#a
#clerical = reference
levels(CPS85$sector)
## [1] "clerical" "const"    "manag"    "manuf"    "other"    "prof"
## [7] "sales"    "service"
```

b:

most = sectormanag, least = sectorsales

c:

For fixed experience, job sector, and marital status, wages increase by 75 cents per hour (on average) for every extra year of education.

Least Squares Estimation

```
#Load the data:
suppressPackageStartupMessages(library(mosaic))
data(Galton)
```

Exercise 12

```
#subject 1 prediction
pred1 <- 39.1104 + 0.3994*75
#subject 1 residual
64.5 - pred1
## [1] -4.565

#subject 2 prediction
pred2 <- 39.1104 + 0.3994*65
#subject 2 residual
67 - pred2
## [1] 1.929
```

Exercise 13

```
#a.
htmodel <- lm(height ~ father, data=Galton)
htmodelResults <-  data.frame(observed=Galton$height,
                              predicted=htmodel$fitted.values, residual=htmodel$residuals)
head(htmodelResults)
##   observed predicted residual
## 1     73.2     70.46    2.738
## 2     69.2     70.46   -1.262
## 3     69.0     70.46   -1.462
## 4     69.0     70.46   -1.462
## 5     73.5     69.26    4.236
```

```
## 6     72.5     69.26    3.236

#b
#residual = observed - predicted

#c
mean(htmodelResults$residual)
## [1] -3.02e-17

#d
#this appears at the top of the model summary
summary(htmodelResults$residual)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -10.300  -2.670  -0.209   0.000   2.630  11.900

#e
#This appears (approximately) as "Residual standard error" in the model summary
sd(htmodelResults$residual)
## [1] 3.444
#a more exact calculation
dim(Galton)
## [1] 898   6
sqrt(sum(htmodelResults$residual^2)/(898-2))
## [1] 3.446
```