Homework 1

Peter Sullivan

What is the association between education and earnings? Using the 1990 dataset entitled "Work, Family, and Well-being in the United States" (https://github.com/avehtari/ROS-Examples/tree/master/Earnings/data, please do the following in this markdown document:

1) read in data found in earnings.csv;
2) graph the data and add a fitted line;
3) fit a linear regression of earnings with education as a predictor;
4) explain what each of the following represents and how it was calculated (see Lab 1c as a reference);

a) b1hat for education
b) b0hat
c) yhat
d) uhat; and

5) interpret the estimated coefficient for education on earnings as well as the R-Squared.

1. Reading in The DAtA

```
urlfile <- "https://raw.githubusercontent.com/avehtari/ROS-Examples/master/Earnings/data/earnings.csv"

data <- read_csv(urlfile, na = c("",NA))
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   height = col_double(),
##   weight = col_double(),
##   male = col_double(),
##   earn = col_double(),
##   earnk = col_double(),
##   ethnicity = col_character(),
##   education = col_double(),
##   mother_education = col_double(),
##   father_education = col_double(),
##   walk = col_double(),
##   exercise = col_double(),
##   smokenow = col_double(),
##   tense = col_double(),
##   angry = col_double(),
##   age = col_double()
## )
```

```
sapply(data, function(Count) sum(is.na(Count)))
```

```
##           height           weight             male             earn
##                0               27                0                0
##            earnk        ethnicity        education mother_education
##                0                0                2              244
## father_education             walk         exercise         smokenow
##              295                0                0                1
##            tense            angry              age
##                1                1                0
```

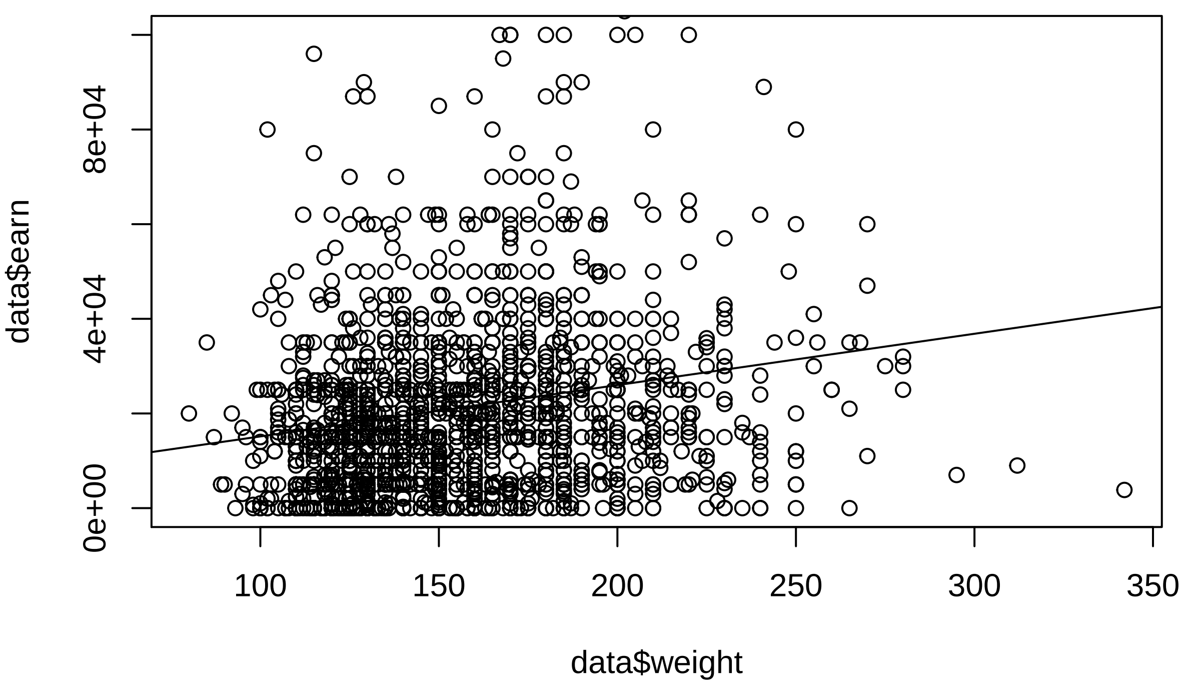
2. Graph the Data and Add a Fitted Line

I’ve decided to look into the correlation between weight and earnings. Weight will be on the x axis, and earnings on the Y.

```
#colnames(data)

fit <- lm(earn ~ weight, data = data)

summary(fit)
```

```
##
## Call:
## lm(formula = earn ~ weight, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -37559 -13655  -3401   6717 375633
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4298.29    2441.98   1.760   0.0786 .
## weight        108.48      15.25   7.112 1.65e-12 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 22330 on 1787 degrees of freedom
##   (27 observations deleted due to missingness)
## Multiple R-squared:  0.02752,    Adjusted R-squared:  0.02698
## F-statistic: 50.58 on 1 and 1787 DF,  p-value: 1.65e-12
```

```
plot(data$weight,data$earn, ylim = c(0,1E5))+abline(fit)
```



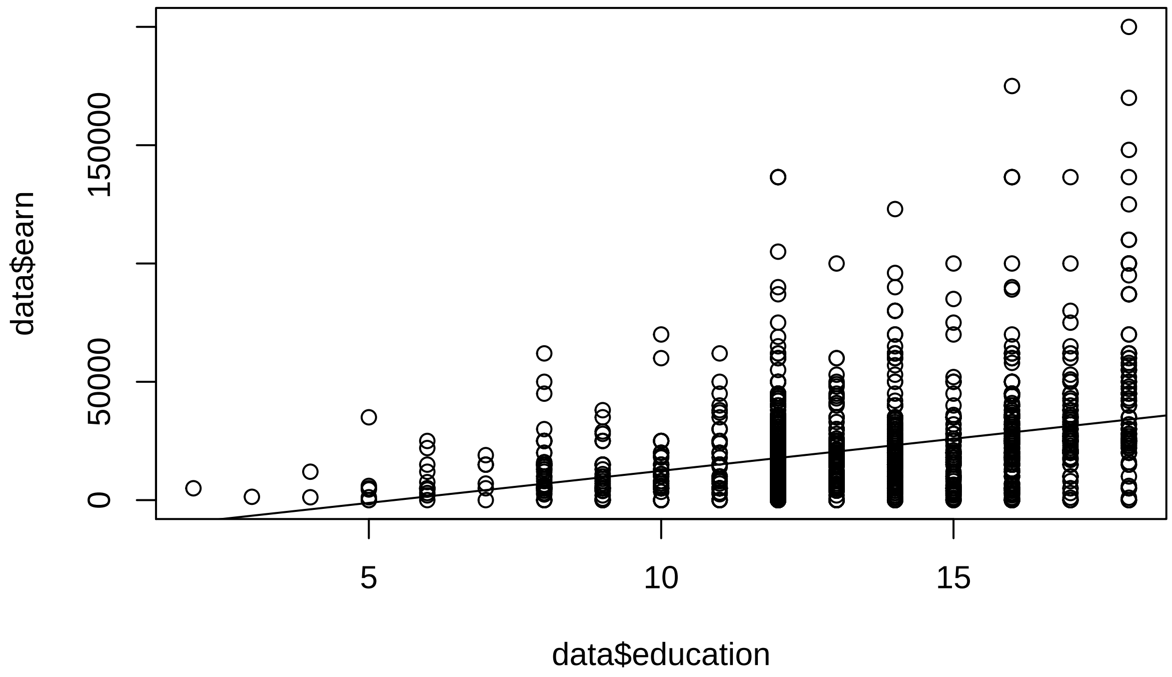
```
## integer(0)
```

3. fit linear regression with education as a indicator

```
#colnames(data)

fit <- lm(earn ~ education, data = data)

summary(fit)
```

```
##
## Call:
## lm(formula = earn ~ education, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -34051 -12373  -3212   7207 382207
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -14724.1     2657.0  -5.542 3.43e-08 ***
## education     2709.7      197.1  13.748  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21460 on 1812 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.09445,    Adjusted R-squared:  0.09395
## F-statistic:   189 on 1 and 1812 DF,  p-value: < 2.2e-16
```

```
plot(data$education,data$earn, ylim = c(0,2E5))+abline(fit)
```



```
## integer(0)
```

4 Explainations:

explain what each of the following represents and how it was calculated (see Lab 1c as a reference);

a) b1hat for education

B1 hat = 2709.7

For every 1 year in education, earnings go up 2709.7 dollars, This is the slop of the regression line.

b) b0hat

B0 hat = -14,724.1

If you had 0 years of education, you would have earned $-14,724. This is the Intercept of the regression line.

c)

yhat = -14,724 + 2709 *X Yhat is the expected value of y given x using our regression model. This is the fitted model using the lm regression.

d) uhat

Uhat represents the residuals from the actual data to the residuals. It can be seen as the reported error from the model per data point. Uhat is the variance from each point the line of best fit. The mean of uhat is basically zero, as we would expect.

```
uhat <- resid(fit)
mean(uhat)
```

```
## [1] -8.319833e-13
```

E: R^2

Multiple R-squared: 0.09445

9.45 % of the variance in the model is explained by the line of best fit. This is a poor model, and I would not recommend using it.

Looking into other variables

```
fit1 <- stan_glm(data$earn~data$education)
```

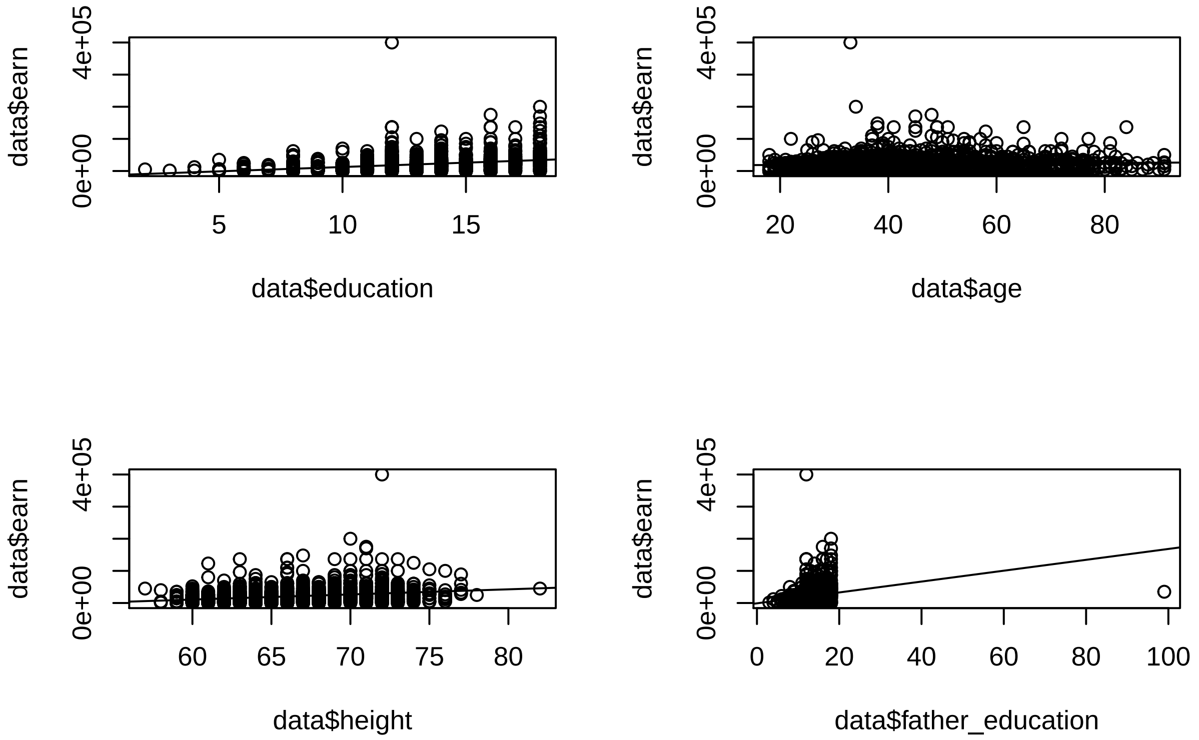
```
## Warning: Omitting the 'data' argument is not recommended and may not be allowed
## in future versions of rstanarm. Some post-estimation functions (in particular
## 'update', 'loo', 'kfold') are not guaranteed to work properly unless 'data' is
## specified as a data frame.
```

```
summary(fit1)

fit2 <- lm(data$earn~ data$age)
summary(fit2)



fit3 <- lm(data$earn~ data$height)
summary(fit3)




fit4 <- lm(data$earn~ data$father_education)
summary(fit4)
```

```
## integer(0)
```

```
## integer(0)
```

```
## integer(0)
```



```
## integer(0)
```

Looping through the models

```
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 4.0.5
```

```
data
```

```
## # A tibble: 1,816 x 15
##    height weight  male  earn earnk ethnicity education mother_education
##     <dbl>  <dbl> <dbl> <dbl> <dbl> <chr>         <dbl>            <dbl>
##  1     74    210     1 50000    50 White            16               16
##  2     66    125     0 60000    60 White            16               16
##  3     64    126     0 30000    30 White           16                16
##  4     65    200     0 25000    25 White           17                17
##  5     63    110     0 50000    50 Other           16                16
##  6     68    165     0 62000    62 Black           18                18
##  7     63    190     0 51000    51 White           17                17
##  8     64    125     0  9000     9 White           15                15
##  9     62    200     0 29000    29 White           12                12
## 10     73    230     1 32000    32 White           17                17
## # ... with 1,806 more rows, and 7 more variables: father_education <dbl>,
## #   walk <dbl>, exercise <dbl>, smokenow <dbl>, tense <dbl>, angry <dbl>,
## #   age <dbl>
```

```
#corr_data <- data[,c(1,2,4,5,7,9:15)]
#var(corr_data, na.rm = FALSE)
#kable(round(cor(corr_data),2))

columns <- as.list(colnames(data))



models <- lapply(paste("earn ~", columns), as.formula)
models
```

```
## [[1]]
## earn ~ height
## <environment: 0x0000000019789c10>
##
## [[2]]
## earn ~ weight
## <environment: 0x0000000019789c10>
##
## [[3]]
## earn ~ male
## <environment: 0x0000000019789c10>
##
## [[4]]
## earn ~ earn
## <environment: 0x0000000019789c10>
##
## [[5]]
## earn ~ earnk
## <environment: 0x0000000019789c10>
##
## [[6]]
## earn ~ ethnicity
## <environment: 0x0000000019789c10>
##
## [[7]]
## earn ~ education
## <environment: 0x0000000019789c10>
##
## [[8]]
## earn ~ mother_education
## <environment: 0x0000000019789c10>
```

```
##
## [[9]]
## earn ~ father_education
## <environment: 0x0000000019789c10>
##
## [[10]]
## earn ~ walk
## <environment: 0x0000000019789c10>
##
## [[11]]
## earn ~ exercise
## <environment: 0x0000000019789c10>
##
## [[12]]
## earn ~ smokenow
## <environment: 0x0000000019789c10>
##
## [[13]]
## earn ~ tense
## <environment: 0x0000000019789c10>
##
## [[14]]
## earn ~ angry
## <environment: 0x0000000019789c10>
##
## [[15]]
## earn ~ age
## <environment: 0x0000000019789c10>
```

```
for (model in models){
  fit <- lm(model, data = data)
  x <-summary(fit)
  print(paste(format(model),"R^2 value: ",round(x$r.squared,3)*100,"%"))
}
```

```
## [1] "earn ~ height R^2 value:  7.4 %"
## [1] "earn ~ weight R^2 value:  2.8 %"
## [1] "earn ~ male R^2 value:  9.4 %"
```

```
## Warning in model.matrix.default(mt, mf, contrasts): the response appeared on the
## right-hand side and was dropped
```

```
## Warning in model.matrix.default(mt, mf, contrasts): problem with term 1 in
## model.matrix: no columns are assigned
```

```
## [1] "earn ~ earn R^2 value:  0 %"
## [1] "earn ~ earnk R^2 value:  100 %"
## [1] "earn ~ ethnicity R^2 value:  0.8 %"
## [1] "earn ~ education R^2 value:  9.4 %"
## [1] "earn ~ mother_education R^2 value:  5.7 %"
## [1] "earn ~ father_education R^2 value:  5.4 %"
## [1] "earn ~ walk R^2 value:  0.2 %"
## [1] "earn ~ exercise R^2 value:  1 %"
## [1] "earn ~ smokenow R^2 value:  0.1 %"
## [1] "earn ~ tense R^2 value:  0.5 %"
## [1] "earn ~ angry R^2 value:  0.5 %"
## [1] "earn ~ age R^2 value:  0.6 %"
```