Predicting Numerical Variables with Linear Regression

Ken Wood

7/27/2024

In Part Three of the course project, you will use linear regression to analyze a data set from the Opportunity Insights project. The variables you will work with in this part of the course project are student_income and parent_income. Both student and parent income are log transformed for this analysis.

You will use linear regression to determine whether the median income of students who attended a particular school is associated with the median income of parents whose children attended that school. Then, you will build a prediction rule for student_income based on parent_income.

To do this, you will quantify the relationship between parent median income and child median income by - visualizing the data, - building a simple linear prediction rule, and - measuring the uncertainty around the prediction rule.

Ultimately, you will use this analysis to test the hypothesis that children who attended schools in which parents had a higher median income also have a higher median income when they are adults.

To begin, run the following code chunk to load the data set, then answer the questions below.

```
# eCornell Hex Codes:
crimson = '#b31b1b'   # crimson
lightGray = '#cecece' # lightGray
darkGray = '#606366'  # darkGray
skyBlue = '#92b2c4'   # skyblue
gold = '#fbb040'      # gold
ecBlack = '#393f47'   # ecBlack

school = read.csv('mrc_table2.csv', header = TRUE, check.names = FALSE)
school = school[,names(school) %in%
                  c('name', 'type', 'tier', 'tier_name', 'mr_kq5_pq1',
                    'par_median', 'k_median')]
names(school)[5:7] <- c("mobility_rate", "parent_income", "student_income")
school$parent_income <- log(school$parent_income)
school$student_income <- log(school$student_income)
```

Step 1

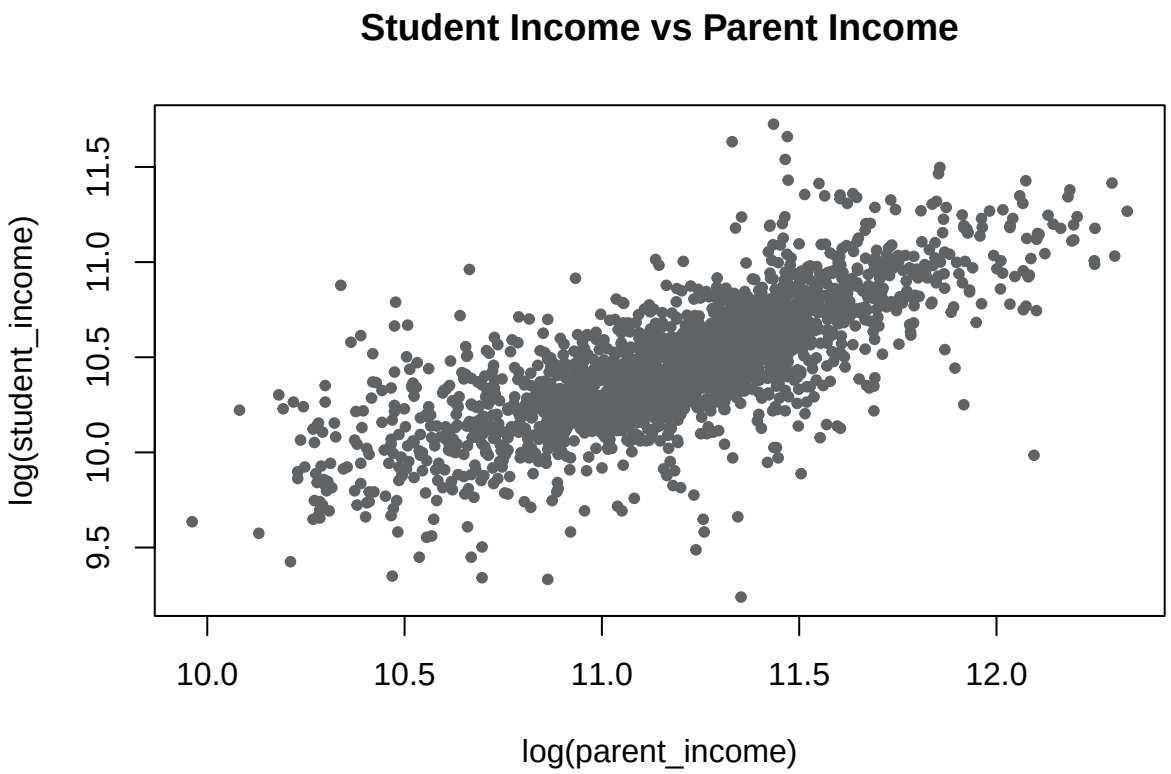
In this analysis, which variable is the predictor variable, and which variable is the response variable?

We want to build a prediction rule for student_income based on parent_income. Therefore, **parent income is the predictor variable** and **student income is the response variable**.

Step 2

Make a scatterplot of student_income vs. parent_income, making sure to plot the predictor variable on the x-axis and the response variable on the y-axis. What is the nature of the association between these variables (positive/negative/none)? Quantify the association with a Pearson correlation.

```
# Create plot of observed data:
plot(school$parent_income, school$student_income,
pch = 20, col = darkGray,
xlab = 'log(parent_income)', ylab = 'log(student_income)',
main = 'Student Income vs Parent Income')
```



We see from the plot that there is a positive correlation between student income and parent income.

```
# Pearson correlation
pearson_cor = cor(school$parent_income, school$student_income, method = c("pearson"))
pearson_cor
```

```
## [1] 0.743423
```

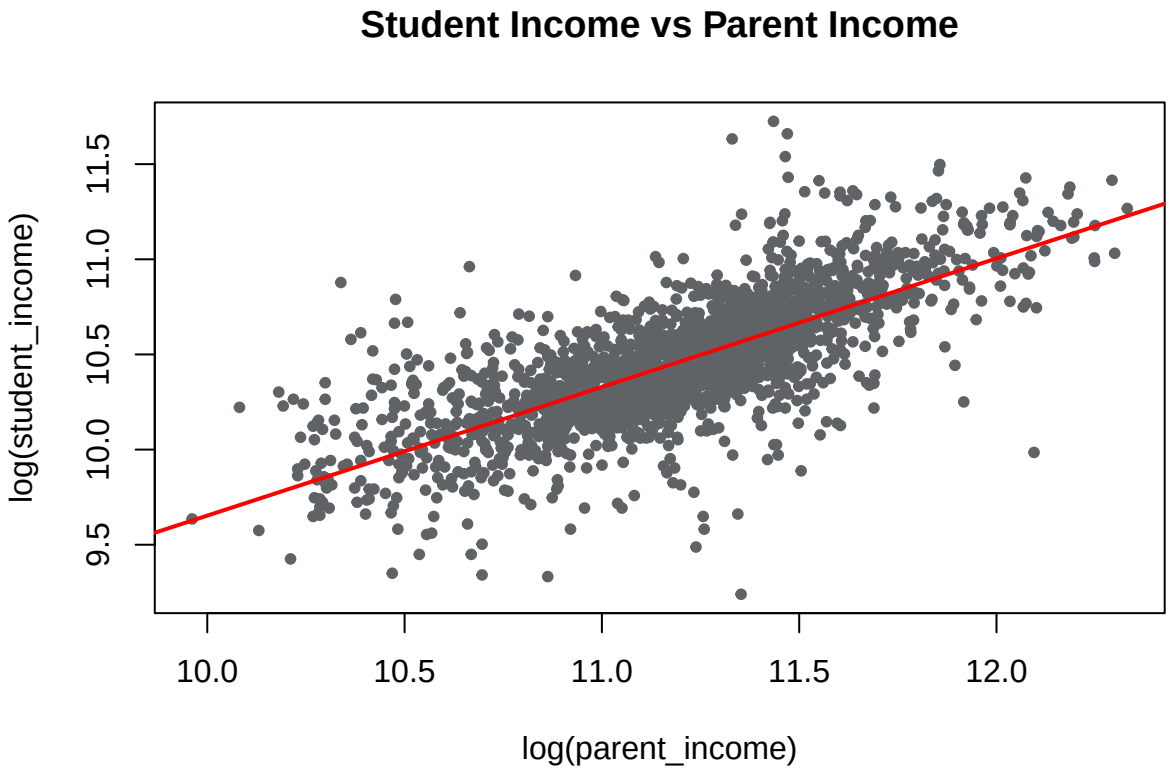
Step 3

Build a linear prediction rule for these data by using the lm() function to fit a regression line through the scatterplot, then add it to your plot. Store your prediction rule in the object fit.studentincome so you can plot it later.

```
fit.studentincome <- lm(student_income~parent_income,data=school)

# Create plot of observed data:
plot(school$parent_income, school$student_income,
pch = 20, col = darkGray,
xlab = 'log(parent_income)', ylab = 'log(student_income)',
main = 'Student Income vs Parent Income')

abline(fit.studentincome,col = "red", lwd = 2)
```



Step 4

Calculate the intercept, slope, and MSE of your regression line. Interpret the meaning of the slope in the context of the hypothesis you're testing.

```
fit.studentincome$coefficients[1]   # Intercept
```

```
## (Intercept)
##    2.896352
```

```
fit.studentincome$coefficients[2]   # Slope
```

```
## parent_income
##     0.6756683
```

```
mean(fit.studentincome$residuals^2) # MSE
```

```
## [1] 0.04774672
```

The slope of the line is 0.68 which tells us that there is a positive, linear relationship between student_income and parent_income.

Step 5

The following code chunk creates 10,000 bootstrapped data sets, fits a regression line to each, stores the value of the correlation and regression coefficients from each bootstrapped data set, and plots the confidence band around the prediction rule.

Run this code chunk, then write the code to calculate and plot the 95% confidence intervals around: - the correlation you calculated in Step 2, and - the slope of the linear prediction rule you built in Step 4.

```
set.seed(1)
B = 10000
```

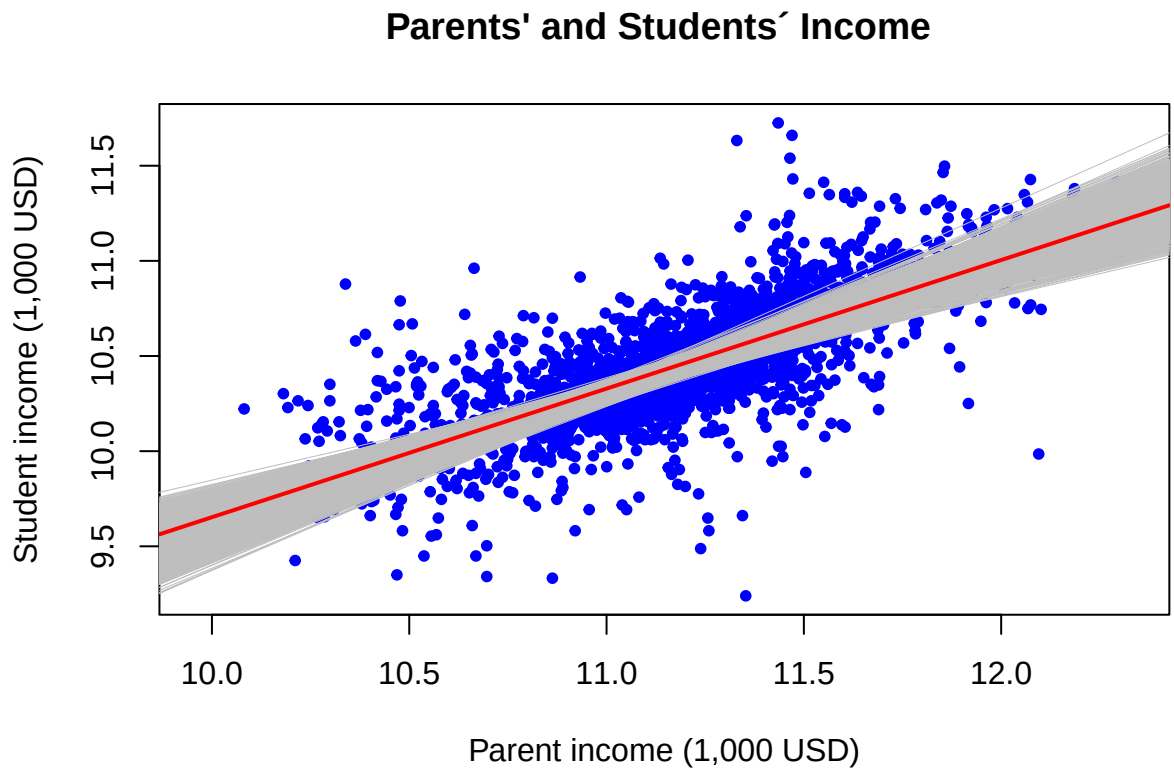
```r
corr.boot = rep(0, B)
a.boot = rep(0,B)
b.boot = rep(0, B)

plot(school$parent_income, school$student_income,
     pch = 20, col = "blue",
     xlab = "Parent income (1,000 USD)", ylab = "Student income (1,000 USD)",
     main = "Parents' and Students' Income ")

for (b in 1:B){
  boot.id = sample(98, replace = TRUE)
  school.boot = school[boot.id,]

  corr.boot[b] = cor(school.boot$parent_income, school.boot$student_income)
  fit.boot <- lm(student_income ~ parent_income, data = school.boot)
  a.boot[b] = fit.boot$coefficients[1]
  b.boot[b] = fit.boot$coefficients[2]

  abline(fit.boot, lwd = 0.5, col = "gray")
}
abline(fit.studentincome, col = "red", lwd = 2)
```

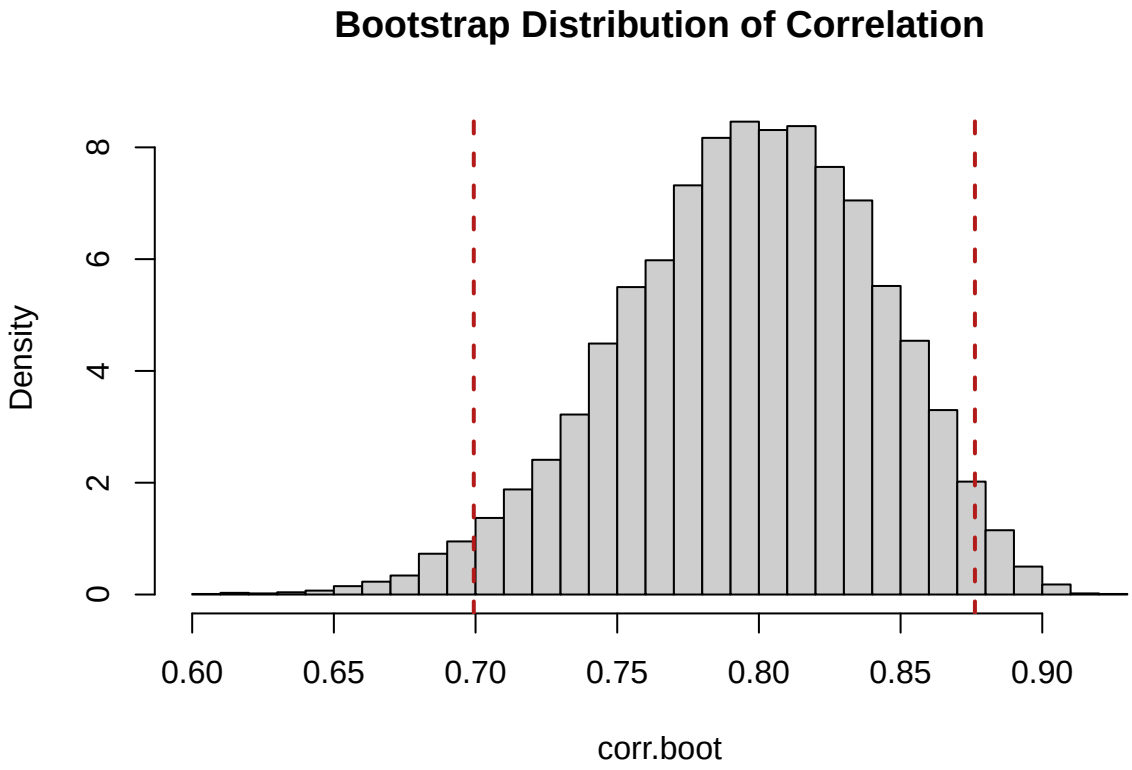


Step 6

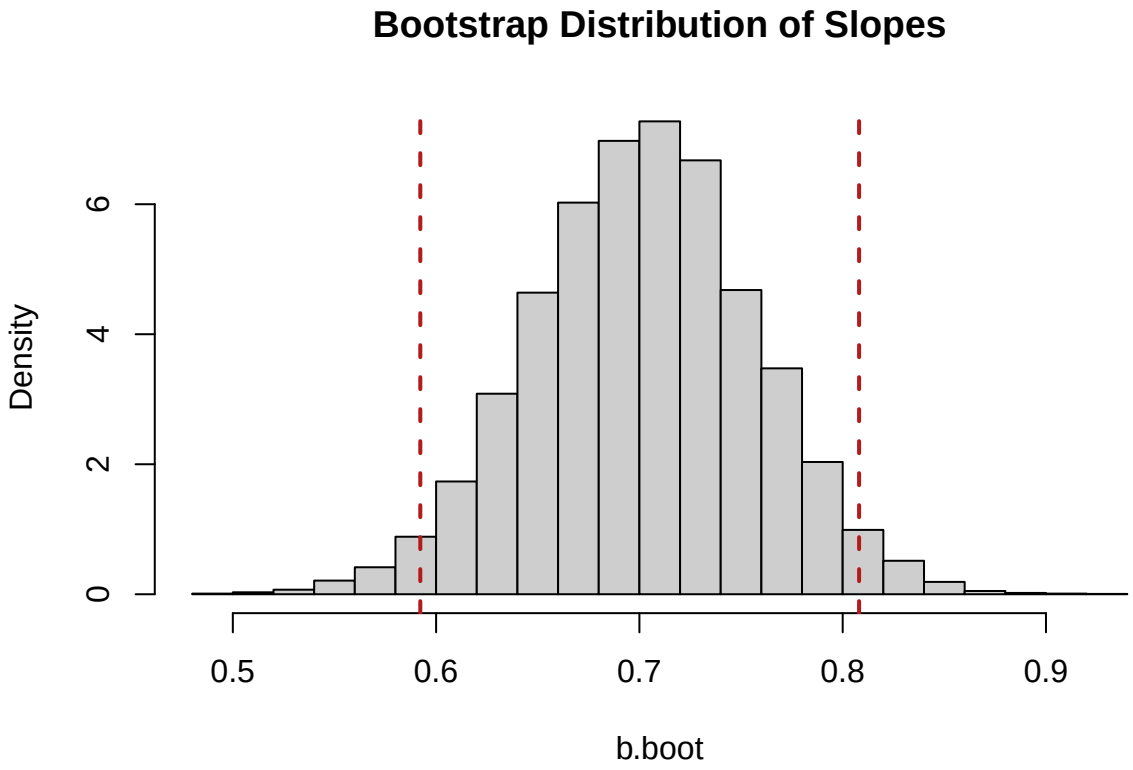
Use the 95% confidence intervals around the correlation coefficient and slope to test the hypothesis that student income is positively correlated with parent income. Briefly explain your answer.

```r
# Create histogram of correlation coefficients (r):
hist(corr.boot, breaks = 30, freq = FALSE, col = lightGray,
```

```
main = 'Bootstrap Distribution of Correlation')
# Add the 95% CI to the histogram:
abline(v = quantile(corr.boot, probs = c(0.025, 0.975)), lty = 2, col = crimson, lwd = 2)
```




```
# Create a histogram of slopes (b):
hist(b.boot, breaks = 30, freq = FALSE, col = lightGray,
main = 'Bootstrap Distribution of Slopes')
# Add the 95% CI to the histogram:
abline(v = quantile(b.boot, probs = c(0.025, 0.975)), lty = 2, col = crimson, lwd = 2)
```



Since both the correlation coefficient = 0.74 and calculated = 0.68 are located within their respective 95% confidence intervals, we can say that we are 95% confident that there is no statistically significant difference between these sample population statistics vs. the total population.

This is the end of Part Three of the course project.

Remember to save and submit your work!