Improving Prediction Rules with Multiple Linear Regression

Ken Wood

7/28/2024

In Part Four of the course project, you will analyze a data set from the Opportunity Insights project and compare two different prediction rules you could use to predict the median income of students when they become adults. You will use this comparison to determine whether the positive relationship you found between the median income of students who attended a particular school is based only on the median income of parents whose children attend that school, or if it is also based on the tier of the school.

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

Use linear regression to build a linear prediction rule that predicts student_income based on tier and parent_income.

```
lr_model1 = lm(student_income~tier+parent_income, data=school)
lr_model1
```

```
##
## Call:
## lm(formula = student_income ~ tier + parent_income, data = school)
##
## Coefficients:
##   (Intercept)           tier  parent_income
##        6.3797        -0.0680         0.4078
```

Step 2

Use the prediction rule you built in Step 1 to describe how student_income would change on average with college tier if parent_income were held constant.

Since the `tier` coefficient is -0.068, student_income would decrease by 6.8% for every positive unit increase of college tier.

Step 3

Describe the influence that parent_income would have on student_income if tier were held constant, based on the prediction rule you built in Question 1.

Since the `parent_income` coefficient is 0.4078, student_income would increase by 40.8% for every positive unit increase of parent_income.

Step 4

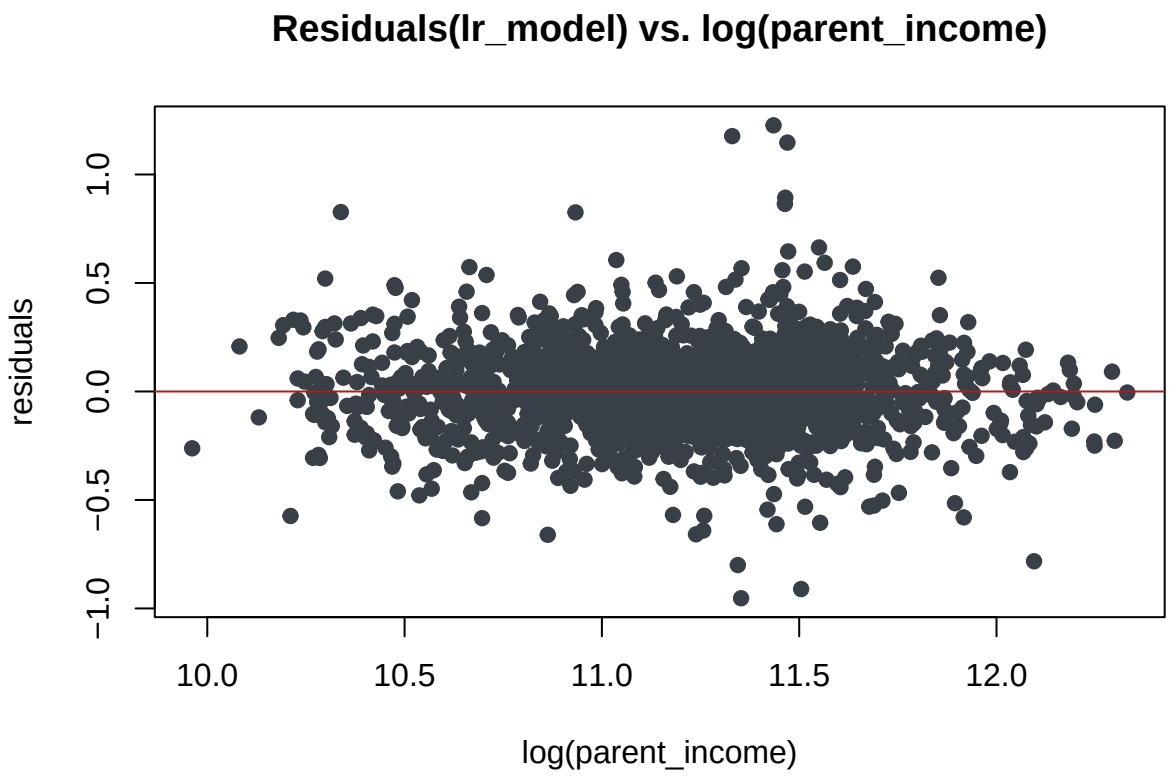
Use your prediction rule to predict the student income based on the tier and parent income for the following four schools: - School A: tier 4, parent_income = 11 - School B: tier 8, parent_income = 11 - School C: tier 4, parent_income = 11.5 - School D: tier 8, parent_income = 11.5

```
school_a = 6.3797 - (0.068 * 4) + (0.4078 * 11)
school_b = 6.3797 - (0.068 * 8) + (0.4078 * 11)
school_c = 6.3797 - (0.068 * 4) + (0.4078 * 11.5)
school_d = 6.3797 - (0.068 * 8) + (0.4078 * 11.5)
```

Step 5

Plot the residuals of the fitted prediction rule vs the parent_income variable. Recall that the parent_income variable was log transformed. Based on the plot of residuals, do you think that this transformation was appropriate? Briefly explain your reasoning.

```
plot(school$parent_income, lr_model1$residuals, col = ecBlack,
     pch = 19, xlab = 'log(parent_income)', ylab = 'residuals',
     main = 'Residuals(lr_model) vs. log(parent_income)')
abline(h = 0, col = crimson)
```



An examination of the residuals plot shows essentially an even distribution about the horizontal zero line. Therefore, the log transformation of parent_income was appropriate.

Step 6

Build a prediction rule for student income with a simple linear regression that predicts student income based only on parent income.

```
lr_model2 = lm(student_income~parent_income, data=school)
lr_model2
```

```
##
## Call:
## lm(formula = student_income ~ parent_income, data = school)
##
## Coefficients:
##   (Intercept)  parent_income
##        2.8964         0.6757
```

Step 7

Calculate the R-squared values for both prediction rules. Use comments to label each calculation.

```
MSE.lr_model1 <- mean(lr_model1$residuals^2)
y = school$student_income
var.y = mean((y - mean(y))^2)
R2.lr_model1 = 1 - MSE.lr_model1/var.y
R2.lr_model1
```

Model 1: lr_model1 = lm(student_income~tier+parent_income, data=school)

```
## [1] 0.6867449
```

```
MSE.lr_model2 <- mean(lr_model2$residuals^2)
y = school$student_income
var.y = mean((y - mean(y))^2)
R2.lr_model2 = 1 - MSE.lr_model2/var.y
R2.lr_model2
```

Model 2: lr_model2 = lm(student_income~parent_income, data=school)

```
## [1] 0.5526777
```

Step 8

Based on the R-squared values of the two prediction rules you assessed above, how much of the variation in student_income was explained by parent_income? How much more variation was explained by adding school tier?

Since the R^2 value for model 2 is 0.55, we can say that parent_income can account for 55% of the variation in student_income. When we add the tier feature in model 1, the R^2 value increases to 0.69 which says that adding tier to the model explains another 14% of the variation in student_income.

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

Remember to save and submit your work!