Exploring Relationships in Mexican Migration Project Data

This exercise was created by Dr. Jacqueline Mauro and is based on the following article:

Garip, Filiz. 2012. "Discovering Diverse Mechanisms of Migration: The Mexico–US Stream 1970–2000." *Population and Development Review*, Vol. 38, No. 3, pp. 393-433.

The data come from the **Mexican Migration Project**, a survey of Mexican migrants from 124 communities located in major migrant-sending areas in 21 Mexican states. Each community was surveyed once between 1987 and 2008, during December and January, when migrants to the U.S. are most likely to visit their families in Mexico. In each community, individuals (or proxy respondents for absent individuals) from about 200 randomly selected households were asked to provide demographic and economic information and to state the time of their first and their most recent trip to the United States. The data included here on the proportion of respondents' income sent to Mexico in the form of remittances was simulated by the teaching assistants of CMU's Statistical Reasoning with R course (90-711).

The data set is the file `migration.csv`. Variables in this dataset can be broken down into two categories:

INDIVIDIUAL LEVEL VARIABLES

Name	Description
`year`	Year of respondent's first trip to the U.S.
`age`	Age of respondent
`male`	1 if respondent is male, 0 if respondent is female
`prop_remitted`	Proportion of respondent's income sent to Mexico in form of remittances
`educ`	Years of education: secondary school in Mexico is from years 7 to 12

COMMUNITY LEVEL VARIABLES

Name	Description
`prop_cmig`	Proportion of respondent's community who are also U.S. migrants
`log_npop`	Logged size of respondent's community.
`prop_self`	Proportion of respondent's community who are self-employed
`prop_agri`	Proportion of respondent's community involved in agriculture
`prop_lessminwage`	Proportion of respondent's community who earn less than the U.S. minimum wage

```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
# ggfortify is a new package for this HW
# Run install.packages("ggfortify") if you do not have it installed

#install.packages("ggfortify")

require(ggfortify)
```

```
## Loading required package: ggfortify
```

```r
require(tidyverse)
```

```
## Loading required package: tidyverse
```

```
## -- Attaching packages ---------------------------------------------------------------- tidyverse 1.3
```

```
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
## v purrr   0.3.4
```

```
## -- Conflicts ------------------------------------------------------------------- tidyverse_conflicts
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
# Load data
migration <- read.csv("data/migration.csv")
```

Question 1 [6 pts]

1a

Calculate the mean values for the individual level and community level characteristics in the dataset. Using these, describe the "average migrant."

1b

Do you think this combination of means is a useful description? Why or why not? List two pieces of information it would be most useful to add to your knowledge of the means and why each is important.

Answer 1

#Answer for 1a The average migrant came during the 1985 time period. The migrants aage is around 24.24. The average migrant was a male being 72%. The Proportion of respondent's income sent to Mexico in form of remittances was around 35%. Years of education: secondary school in Mexico is from years 7 to 12 is around 6.79 years.

#Answer for 1b I think this combination of means is useful to know the average outcome for the type of people that are involved in this study. We can also analyse the range of date as well as gaps and outliers that exist in our data, we can also use this information to see if our data is skewed in any particular direction. We can also use the standard deviation to further analyze the data and to figure out the distribution of the data. We can see if our data is clustered ir more spread out for the variables. We will also be able to get a lot more information if we can analyse teh shape of our distribution and look ar the spread of the data.

```
migration%>%
  summarize_at(vars('year', 'age', 'male', 'prop_remitted', 'educ'), funs(mean)) %>%
  gather(variable, value)
```

```
## Warning: `funs()` is deprecated as of dplyr 0.8.0.
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
##
##   # Auto named with `tibble::lst()`:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

```
##        variable        value
## 1          year 1985.8323655
## 2           age   24.2435333
## 3          male    0.7202182
## 4 prop_remitted    0.3561234
## 5          educ    6.7935363
```

```
migration%>%
  summarize_at(vars('prop_cmig','log_npop','prop_self','prop_agri','prop_lessminwage'),funs(mean)) %>%
  gather(variable, value)
```

```
##           variable     value
## 1        prop_cmig 0.1049755
## 2         log_npop 8.9238451
```

```
## 3        prop_self 0.3450283
## 4        prop_agri 0.3743833
## 5 prop_lessminwage 0.1385603
```

Question 2 [8 pts]

2a

Create scatterplots to investigate the relationship between `prop_self` and `prop_agri`, as well as the relationship between `prop_self` and `log_npop`. Briefly interpret these scatter plots and what they imply about self-employed workers. Is knowing that a migrant is from an area where more people are self-employed informative about these two other aspects of their area?

2b

Calculate the linear correlation for all possible pairs of the four community level variables: `prop_self`, `prop_agri`, `prop_lessminwage`, and `log_npop`. Use these correlations to help with your interpretation of the scatter plots. Does adding the information for the `prop_lessminwage` variable add anything to your interpretation?

Answer 2

#Answer for 2a

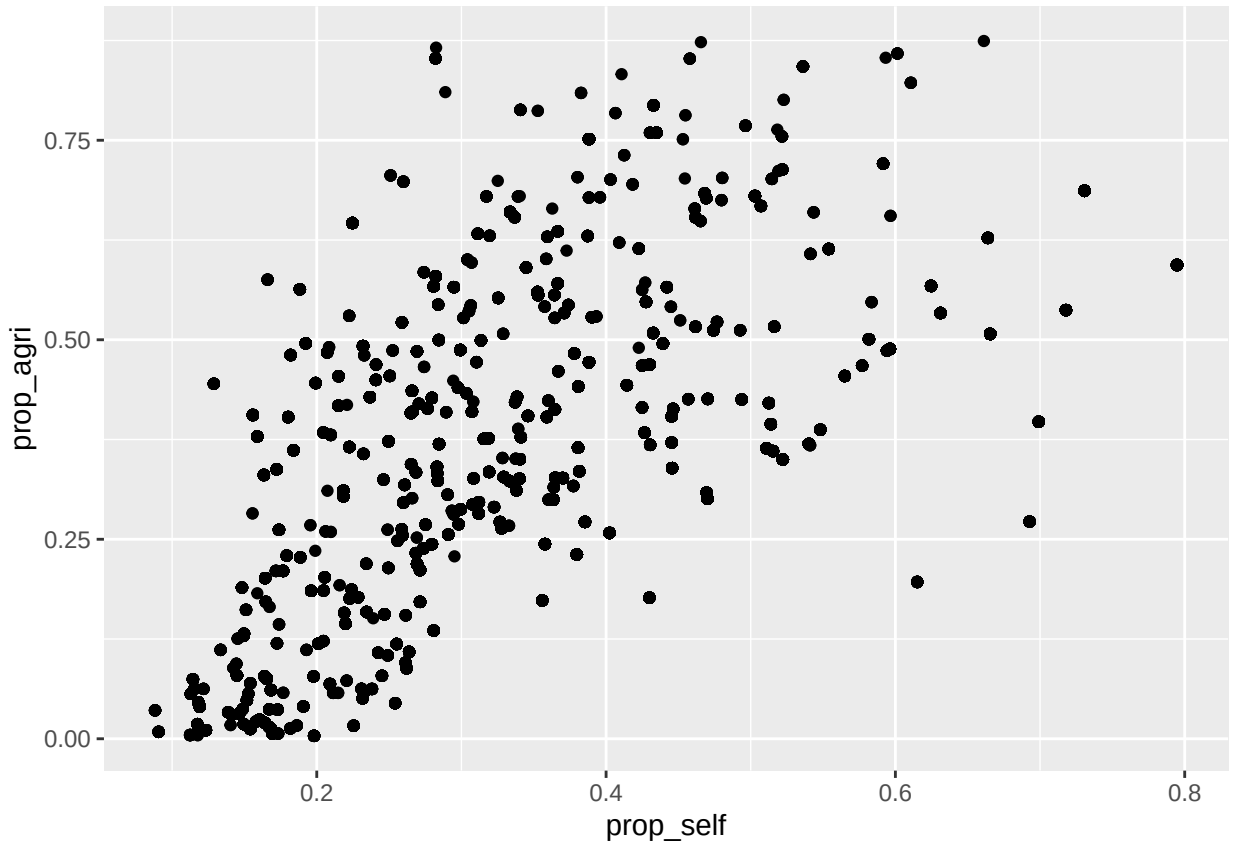
The relation we have observed for the data is positively linearly correlated for the Proportion of respondent's community who are self-employed and Proportion of respondent's community involved in agriculture but we can see that the data is not tightly close together. We can still see that the data is dispersed so the relationship is not very strong. We can also see as our X value is increased , our Y value is also increased.

The relationship we have observed for the data is negatively linearly correlated for the Proportion of respondent's community who are self-employed and Logged size of respondent's community but we can see that the data is not tightly close together. We can still see that the data is dispersed so the relationship is not very strong. The relationship we have observed is that the x increased y decreased.

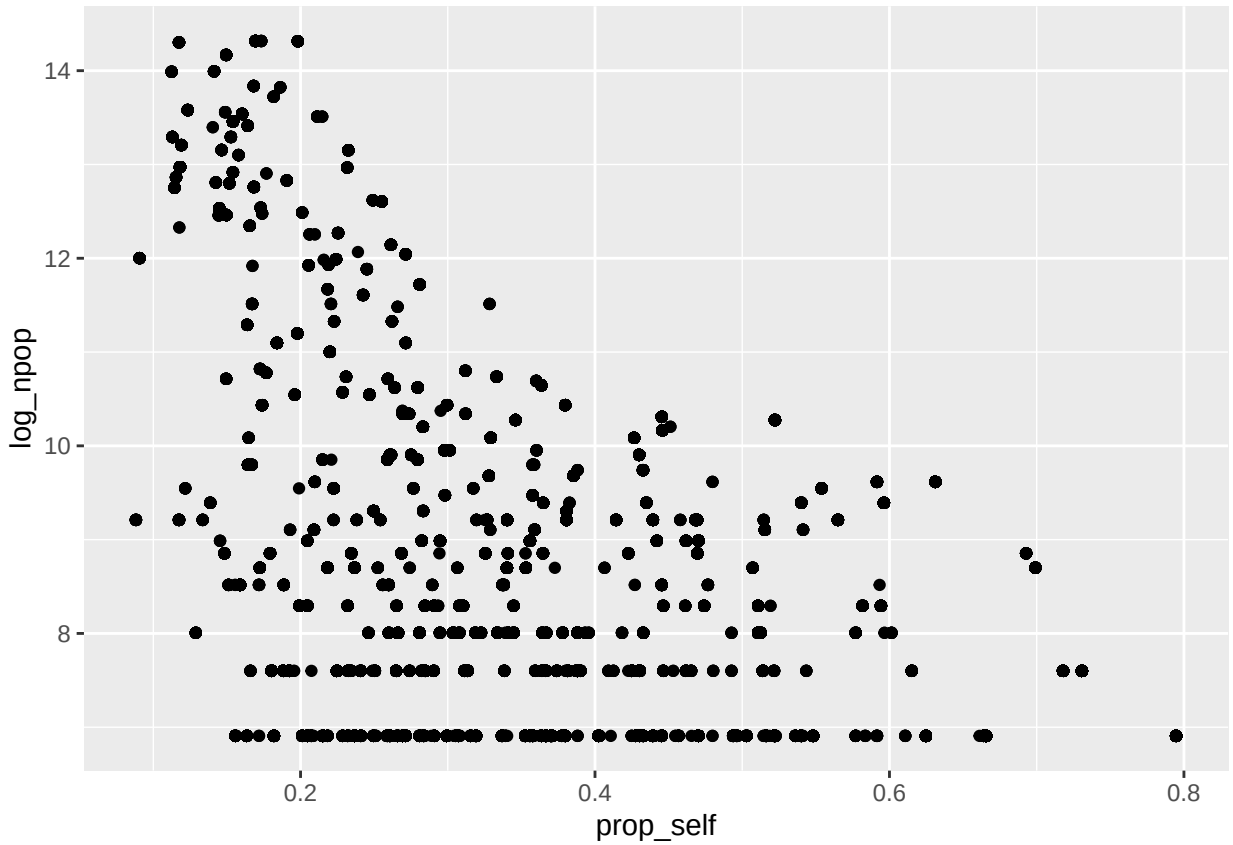
#Answer for 2b

We can see that the relationship of the correlation between the self-employed workers and and agri workers does show us a positive correlation that is .54. Also between the residents that earn less than the USA min wage and self employed workers, we see a a negative correlation at -0.108. We can observe a negative correlation between the log sized correspondents and the self employed workers at -0.4319743. We get a positive correlation at 0.3738637 between the proportion involved in the agriculture and minimum wage less than the average in the USA. We get a negative correlation at -0.6521437 for the respondents in the agriculture group and logged sized respondent community. We also get a negative correlation at -0.05677052 for the logged sized respondent community and proportions of respondents who earn less than the USA min wage. As seen with this, the variables that we see the weakest correlation is the agriculture and log sized respondents.

```
migration %>%
  ggplot(aes(x = prop_self, y = prop_agri)) +
  geom_point()
```




```
migration %>%
  ggplot(aes(x = prop_self, y = log_npop)) +
  geom_point()
```




```
cor(migration$prop_self, migration$prop_agri)
```

```
## [1] 0.5411598
```

```
cor(migration$prop_self, migration$prop_lessminwage)
```

```
## [1] -0.1079667
```

```
cor(migration$prop_self, migration$log_npop)
```

```
## [1] -0.4319743
```

```
cor(migration$prop_agri, migration$prop_lessminwage)
```

```
## [1] 0.3738637
```

```
cor(migration$prop_agri, migration$log_npop)
```

```
## [1] -0.6521437
```

```
cor(migration$prop_lessminwage, migration$log_npop)
```

```
## [1] -0.05677052
```

Question 3 [7 pts]

3a

Check if the relationship between the proportion of people in a migrant's community who are self-employed and the proportion of people working in the agricultural sector in a migrant's community can be usefully modeled by a linear regression.

To do this, regress the proportion of self-employed people in the community (this is the outcome or response variable) on the proportion of people working in agriculture in the community (this is the predictor variable). Create a scatterplot showing the relationship between these two variables and add the estimated regression line to the figure.

3b

Then create a scatterplot with the model residuals on the vertical axis and the predictor (X) values on the horizontal axis.

3c

Assess these figures to determine if a linear regression model is useful for understanding this bivariate relationship. State whether the linearity assumption holds or is violated and describe what about the figures led you to this conclusion.

##Answer 3

#Answer for 3c We see a positive linear association on the relationship between the proportion of self employed migrants and agriculture workers. The RMSE is 12% points away from what is expected based on the the linear relationship with the proportions of respondents that work with agriculture. The residual plot will show us that the mena is close to 0 for all the values of the prediction var. This is also supported with the linearity assumption for the self-employed workers and the agriculture workers holds.

```
regline1 <- lm(prop_self ~ prop_agri, data = migration)
summary(regline1)
```

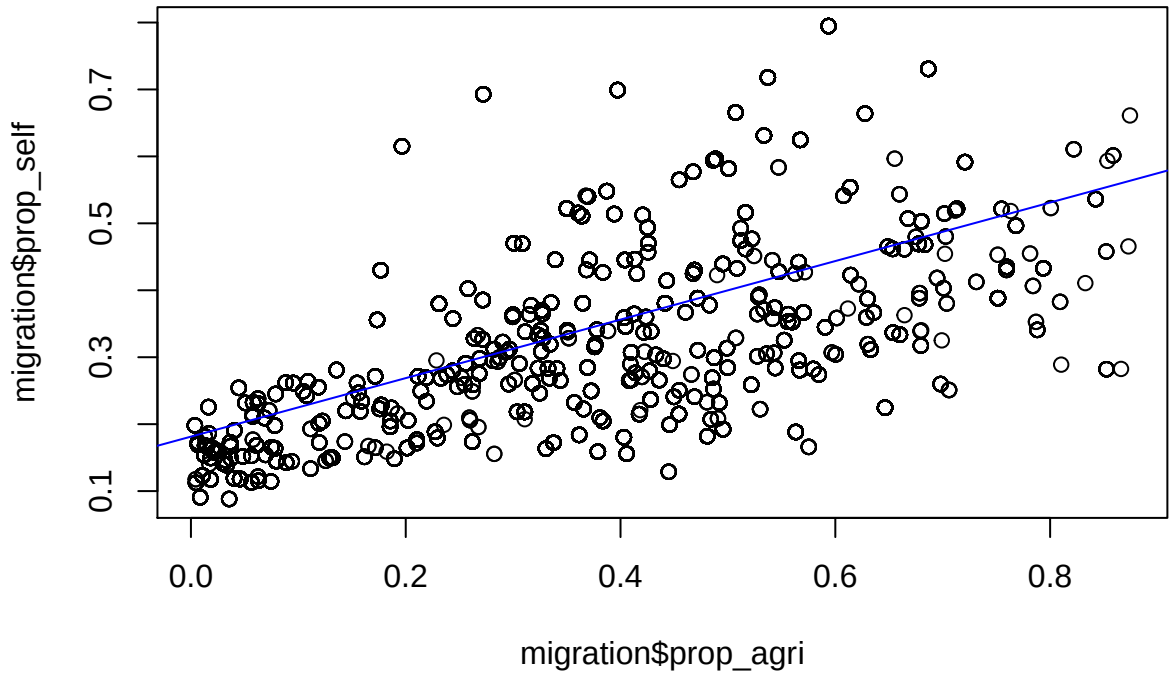
```
##
## Call:
## lm(formula = prop_self ~ prop_agri, data = migration)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.27745 -0.09467 -0.00836  0.06906  0.39246
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.181374   0.002166   83.73   <2e-16 ***
## prop_agri   0.437132   0.005203   84.02   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1238 on 17047 degrees of freedom
## Multiple R-squared:  0.2929, Adjusted R-squared:  0.2928
## F-statistic:  7060 on 1 and 17047 DF,  p-value: < 2.2e-16
```

```
regline1residual <- resid(regline1)
summary(regline1residual)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
```

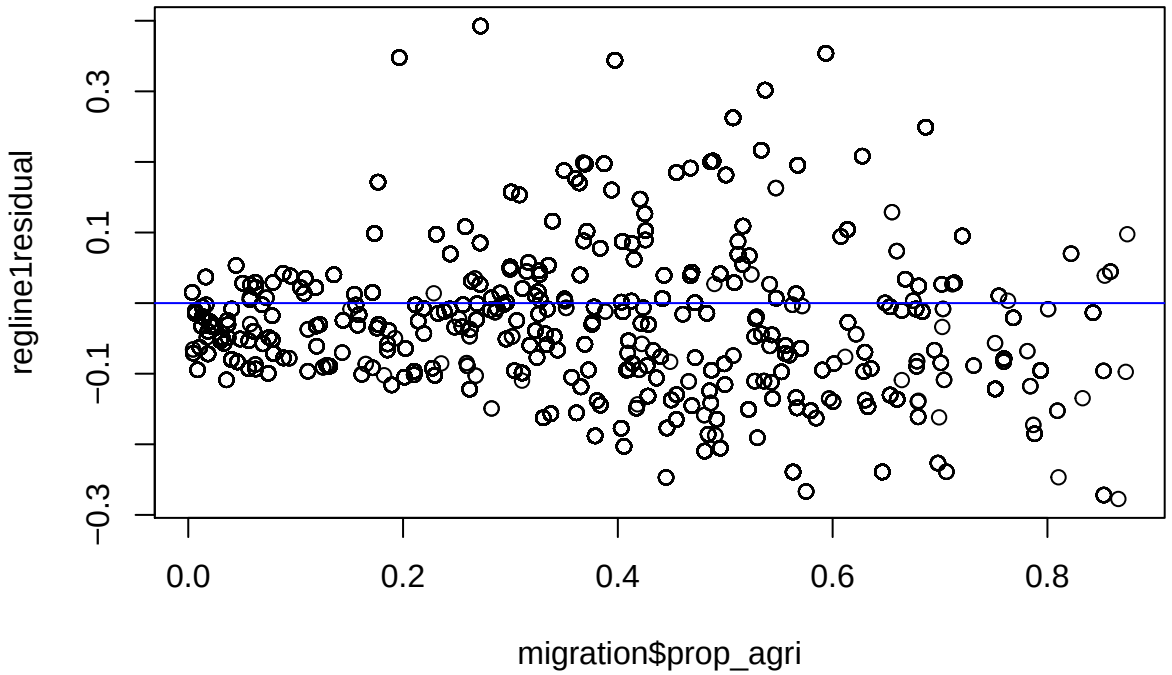
```
## -0.277450 -0.094671 -0.008357  0.000000  0.069059  0.392461
```

```
plot(migration$prop_agri, migration$prop_self) + abline(0.181374, 0.437132, col = "blue")
```




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

```
plot(migration$prop_agri, regline1residual)
abline(h=0, col = "blue")
```



Question 4 [9 pts]

Use the linear regression model you estimated in Question 3. Whether or not you concluded that the linearity assumptin held in the prior question, for the purpose of this question, assume it did hold.

4a

Write out the regression equation and interpret the value of the y-intercept. Is this value practically meaningful? Why or why not?

4b

Interpret the value of the slope coefficient. Describe what this number tells you in words. Describe the slope on a meaningful scale. Interpret the residual (see slide 13 of Lecture 6a for an example of this).

4c

Consider a new respondent to the survey in a community where the proportion of workers involved in agriculture is 0.2. Using the linear regression results, what do you estimate the proportion of self-employed workers to be in this new respondent's community?

4d

State and interpret the value of the RMSE and relate it to your answer to 4c.

4e

State and interpret the R^2 of the model.

Answer 4

#Answer for 4a prop_self = .18 + .43*prop_agri + residual

This value is practically meaningful because this equation shows us the true line, which is the underlying data generating mechanism and this is not an estimate.

#Answer for 4b

We see a positive coefficient which indicates that the value of the independent variable (prop_self) increases and the mean of the dependent variables (prop_agri) is also in a trend to increase.

Residual: The residual number represents how far the proportions of self employed migrant is from the number we expected based on the linear relationship with the number of respondents for the agriculture respondents.

#Answer for 4c prop_self = .18 + .43*prop_agri prop_self = .18 + .43*.2 prop_self = .266

We have use the new value for agriculture respondents, we expect the proportion of the workers to be .266. This does not take into account the RMSE and the data that we got itn the scatter plot in the previous problem. This estimate may not be accurate.

#Answer for 4d

The RMSE is 12.4% points from which we can then conclude that the percentage of self employed migrant workers is 12.4% points off from what is expected based on the linear relationship with the proportion of respondents in agriculture.

#Answer for 4e

The model that we see with R^2 value is the proportion of the variation of prop_self that is also used in calculating the linear relationship with prop_agri. This shows us the information that shows the regression predictions approximates the real data points.

```
lm(formula = prop_self ~ prop_agri, data = migration)
```

```
##
## Call:
## lm(formula = prop_self ~ prop_agri, data = migration)
##
## Coefficients:
## (Intercept)    prop_agri
##      0.1814       0.4371
```

Question 5 [7 pts]

Check if the relationship between the proportion of a migrant's income that is sent back to Mexico in the form of remittances (this is the outcome or response variable) and the proportion of people in a migrant's community who are also migrants (this is the predictor variable) can be appropriately modeled by a linear regression. Follow the same steps as you did in Questions 3 and 4.

5a

Repeat the steps described in Question 3 but for Y = proportion of a migrant's income that is sent back to Mexico in the form of remittances and X = the proportion of people in a migrant's community who are also migrants.

5b

Write out the estimated regression equation and interpret the slope, Y-intercept, and a residual of the estimated linear regression.

5c

Consider a new respondent to the survey in a community where the proportion of people in a migrant's community who are also migrants is 0.15. Using the linear regression results, what do you estimate the proportion of the new respondent's income that is sent back to Mexico in the form of remittances to be? What is the RMSE for this model and how does it relate to this prediction?

5d

State and interpret the R^2 of the model

Answer 5

#Answer for 5b

prop_cmig = .46 - 1.01 * prop_rimmited + residual

residual = .05

#Answer for 5c

prop_cmig = .46 - 1.01 * prop_rimmited prop_cmig = .46 - 1.01 * 0.15 prop_cmig = .3085

The proportion of people in the migrant community, who are also migrants is .15. We can also estimate the proportion of the new respondents income for the people sent back to Mexico in the form of rimmitence would be .3085. The RMSE for this model is .04985, from this we can conclude that our prediction is .05 % points off from using the linear relationship with the proportion of respondents communityt who are also USA migrants.

#Answer for 5d

The model for the proportion for the variance in prop_cmig that is accounted through the linear relationship with prop_rimmtted. We can conclude that the regression prediction, approximates the real data points. The variability for the respondents community who are also USA migrants is taken into account for the proportion of remittance (Multiple R^2 is .29). 29.3% of the variation of the prop_self variable can be shown by the variation in the prop_agri variable. The percentage is low and in order to come up with a conclusion about the association we would need a higher R^2 percentage. (R^2 is .719)

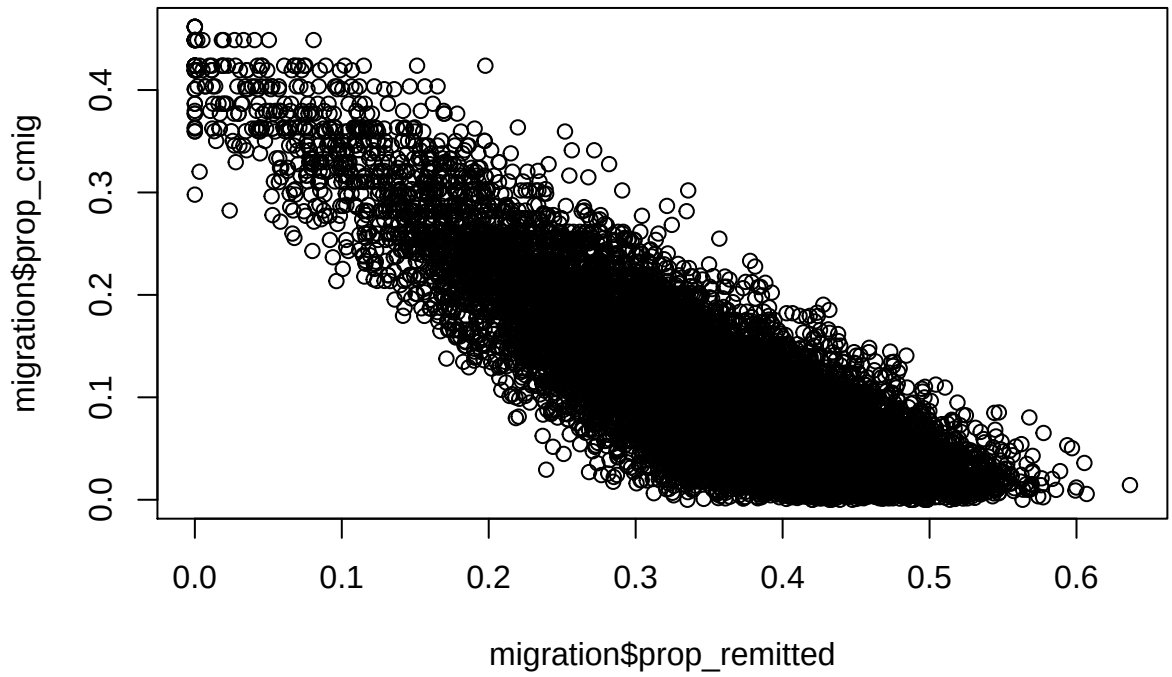
```
regline2 <- lm(prop_remitted ~ prop_cmig, data = migration)
summary(regline2)
```

```
##
## Call:
## lm(formula = prop_remitted ~ prop_cmig, data = migration)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.193438 -0.033725  0.000254  0.033827  0.188654
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.4622090  0.0006355   727.3   <2e-16 ***
## prop_cmig   -1.0105741  0.0048393  -208.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04985 on 17047 degrees of freedom
## Multiple R-squared:  0.719,  Adjusted R-squared:  0.7189
## F-statistic: 4.361e+04 on 1 and 17047 DF,  p-value: < 2.2e-16
```

```
regline2residual <- resid(regline2)
summary(regline2residual)
```

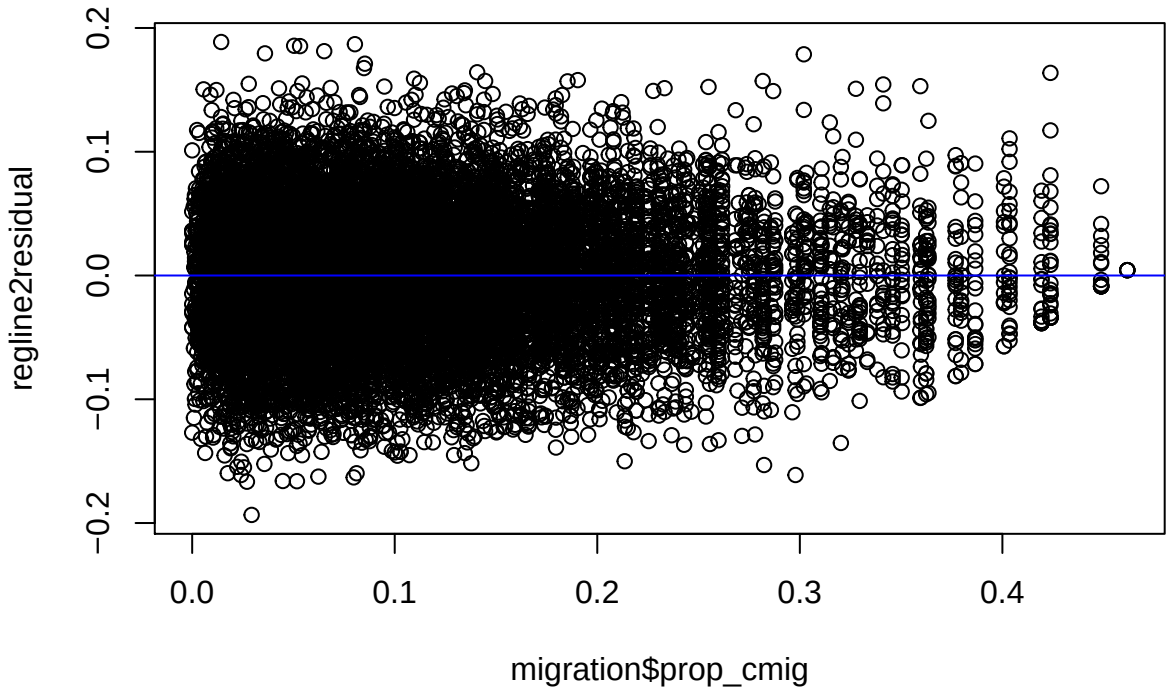
```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -0.1934383 -0.0337255  0.0002544  0.0000000  0.0338274  0.1886539
```

```
plot(migration$prop_remitted, migration$prop_cmig) + abline(-1.0105741, 0.4622090, col = "blue")
```




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

```
plot(migration$prop_cmig, regline2residual) + abline(h=0, col = "blue")
```


regline2residual
0.2
0.1
0.0
−0.1
−0.2
0.0
0.1
0.2
0.3
0.4
migration$prop_cmig

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