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 not male
`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

```r
# Load packages
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

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

migration %>%
  summarize_all(~mean(is.na(.)))
```

```
##   year age male prop_remitted educ prop_cmig log_npop prop_self prop_agri
## 1    0   0    0             0    0         0        0         0         0
##   prop_lessminwage
## 1                0
```

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 1a

```r
# Code for 1a
summary(migration)
```

```
##       year           age             male         prop_remitted   
##  Min.   :1966   Min.   :15.00   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:1979   1st Qu.:18.00   1st Qu.:0.0000   1st Qu.:0.3044  
##  Median :1986   Median :22.00   Median :1.0000   Median :0.3697  
##  Mean   :1986   Mean   :24.24   Mean   :0.7202   Mean   :0.3561  
##  3rd Qu.:1992   3rd Qu.:28.00   3rd Qu.:1.0000   3rd Qu.:0.4213  
##  Max.   :2002   Max.   :65.00   Max.   :1.0000   Max.   :0.6364  
##       educ          prop_cmig          log_npop        prop_self      
##  Min.   : 0.000   Min.   :0.00000   Min.   : 6.908   Min.   :0.08821  
##  1st Qu.: 5.000   1st Qu.:0.04582   1st Qu.: 7.601   1st Qu.:0.23424  
##  Median : 6.000   Median :0.08669   Median : 8.700   Median :0.32665  
##  Mean   : 6.794   Mean   :0.10498   Mean   : 8.924   Mean   :0.34503  
##  3rd Qu.: 9.000   3rd Qu.:0.14228   3rd Qu.: 9.903   3rd Qu.:0.43930  
##  Max.   :25.000   Max.   :0.46166   Max.   :14.316   Max.   :0.79469  
##    prop_agri        prop_lessminwage
##  Min.   :0.003632   Min.   :0.1298  
##  1st Qu.:0.261930   1st Qu.:0.1319  
##  Median :0.372650   Median :0.1373  
##  Mean   :0.374383   Mean   :0.1386  
##  3rd Qu.:0.495250   3rd Qu.:0.1455  
##  Max.   :0.874364   Max.   :0.1565  
```

Text Answer 1a

Before describing the "average migrant", it is relevant to mention that on average 72% of the migrants (migrant himself or the proxy respondents for absent individuals) in this study are male, so the interpretation of results for the 'average migrant' are being done from a male perspective. Now, the average migrant had his first trip to the US in 1986, is of 24.24 years of age, remitted 35.61% of his total income to Mexico, has, on average, 6.8 years of education - which means, he did not go to secondary school (which is from 7-12 years in Mexico), 10.5% of his community is also US migrant, mean of the logged size of average migrant's community is 8.924 (difficult to interpret results as being on log scale and can be better interpreted if on original scale), 35% of his community is self-employed, 37.4% of his community is involved in agriculture, and around 14% of his community is earning less than the minimum wage in the US. ### Text Answer 1b To some extent, this combination of means is giving us an idea about the average migrant; however, interpretation on basis of just mean can be misleading because mean is susceptible to outliers and is not robust. So, we may add some more summary statistics like median, range, IQR, and standard deviation which will make interpretation of data more close to reality. Adding median will give a better idea about each variable in this study as median is more robust in comparison to mean to outliers e.g. mean age will tend towards the outlier age while median age will be a better statistic in this case. Also, if we add range statistic for each variable, then we can get better idea about the overall spread of each variable. Further, standard deviation can give us idea about the spread within data points from the mean. Overall, the more summary statistics we have about the data, the better our understanding gets about that data and thus we get in better position to describe the sample data.

Question 2 [11 pts]

2a (8 points)

Create scatterplots to investigate the bivariate relationship between `prop_self` and `prop_agri`, as well as the bivariate relationship between `prop_self` and `log_npop`. In both figures put `prop_self` on the horizontal axis. 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 (predictive of) these two other aspects of their area?

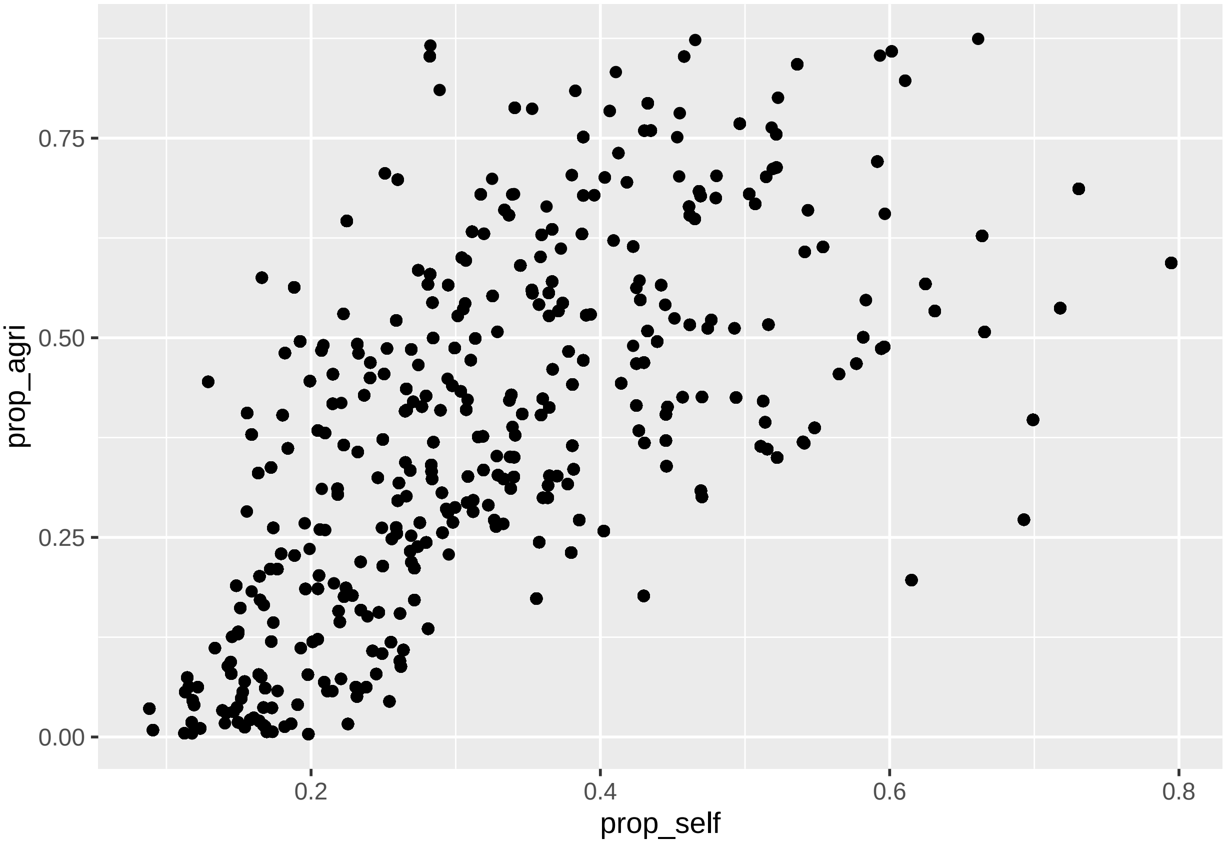
2b (3 points)

Calculate the linear correlation for all possible pairs of the four community level variables: `prop_self`, `prop_agri`, `prop_lessminwage`, and `log_npop`. Which pair has the strongest correlation?

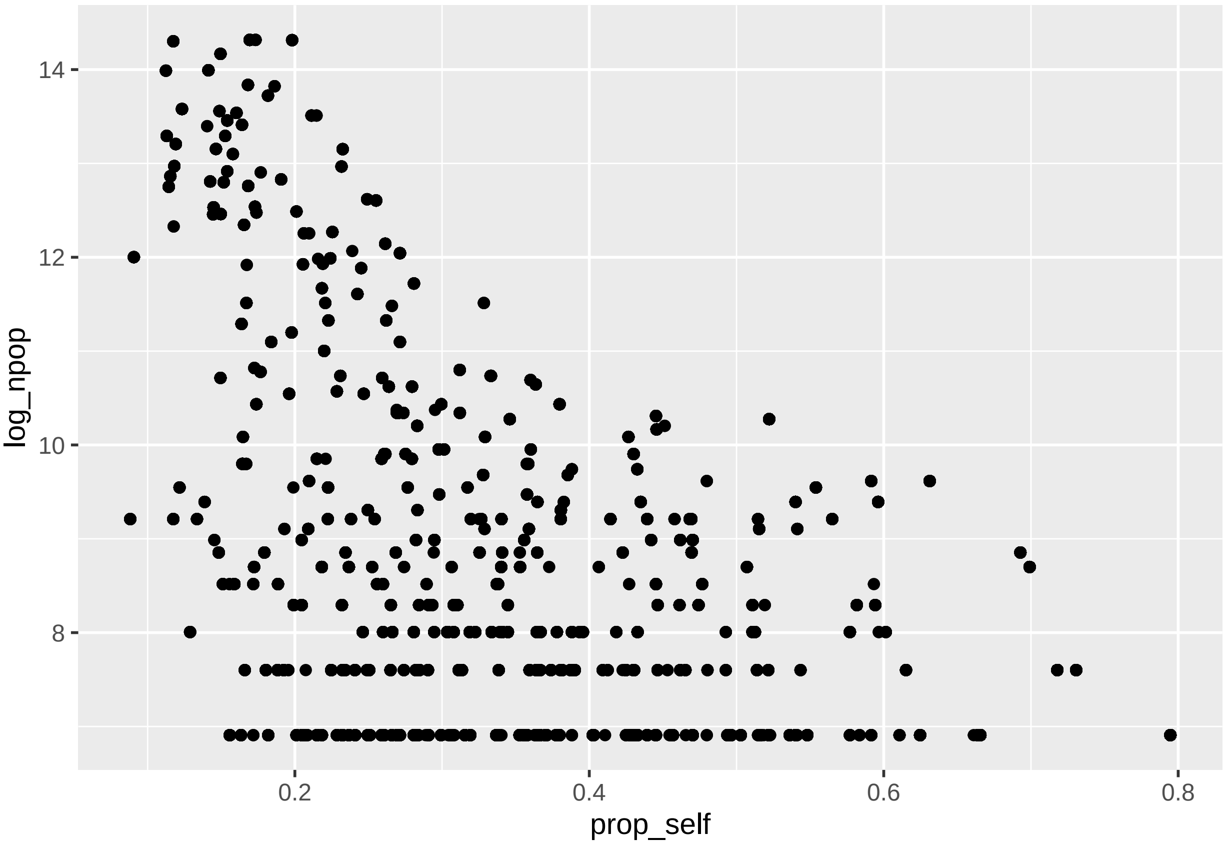
Answer 2

Answer 2a

```r
# Code for 2a
migration %>%
  ggplot(aes(y = prop_agri, x = prop_self)) +
  geom_point()
```



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



### Text Answer 2a We see a medium positive association between the proportions of respondents' community which is self-employed and which is involved in agriculture. Self-employed proportion of the community ranges between 8% to 80% while the proportion of community involved in agriculture ranges between 0% to 88%. Most of the proportion of self-employed community falls below 60% with an outlier of around 80% while there are no distinctively different data points or outliers in proportion of community involved in agriculture and mostly scattered within the mentioned range. There are no clusters of data at some specific points. The figure gives an idea about the medium positive linear association between the proportions of community which is self-employed and which is involved in agriculture. This scatter plot implies that most of the self-employed workers are associated with agriculture.

We see a medium negative association between the proportions of respondents' community which is self-employed and the logged size of respondents' community. Self-employed proportion of the community ranges between 8% to 80% while the logged size of respondents' community ranges between 7 to 14. Most of the proportion of self-employed community falls below 60% with an outlier of around 80% while there are 5 outliers in logged sized of respondent's community while rest of the data points are scattered within the mentioned range. There are no clusters of data at some specific points; however, there are lines of data points at and below 8 logged size community. The figure gives an idea about the medium negative linear association between the proportions of community which is self-employed and logged size of community. Also, on the lower end of logged population below 8, it associates more strongly with the self-employed proportion of population and this strong association is resulting in the lines of data points - which indicates that in small communities more people are self-employed. This scatter plot implies that communities with lesser logged population have greater self-employed proportion. Yes, if we know that a migrant is from an area where more people are self-employed, to some extent, it helps us in predicting that in his community, larger proportion is involved in agriculture and logged size of population of his community is small (but the intensity of the relation of self-employed with these two aspects in not very strong). ### Answer 2b

```
# Code for 2b
migration %>%
  select(prop_self, prop_agri, prop_lessminwage, log_npop) %>%
  cor()
```

```
##                   prop_self  prop_agri prop_lessminwage    log_npop
## prop_self         1.0000000  0.5411598      -0.10796669 -0.43197430
## prop_agri         0.5411598  1.0000000       0.37386371 -0.65214371
## prop_lessminwage -0.1079667  0.3738637       1.00000000 -0.05677052
## log_npop         -0.4319743 -0.6521437      -0.05677052  1.00000000
```

Text Answer 2b

Proportion of population involved in agriculture (prop_agri) and logged size of population (log_npop) pair has the strongest correlation.

## Question 3 [8 pts]

3a (3 points)

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 (2 points)

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

3c (3 points)

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

Answer 3

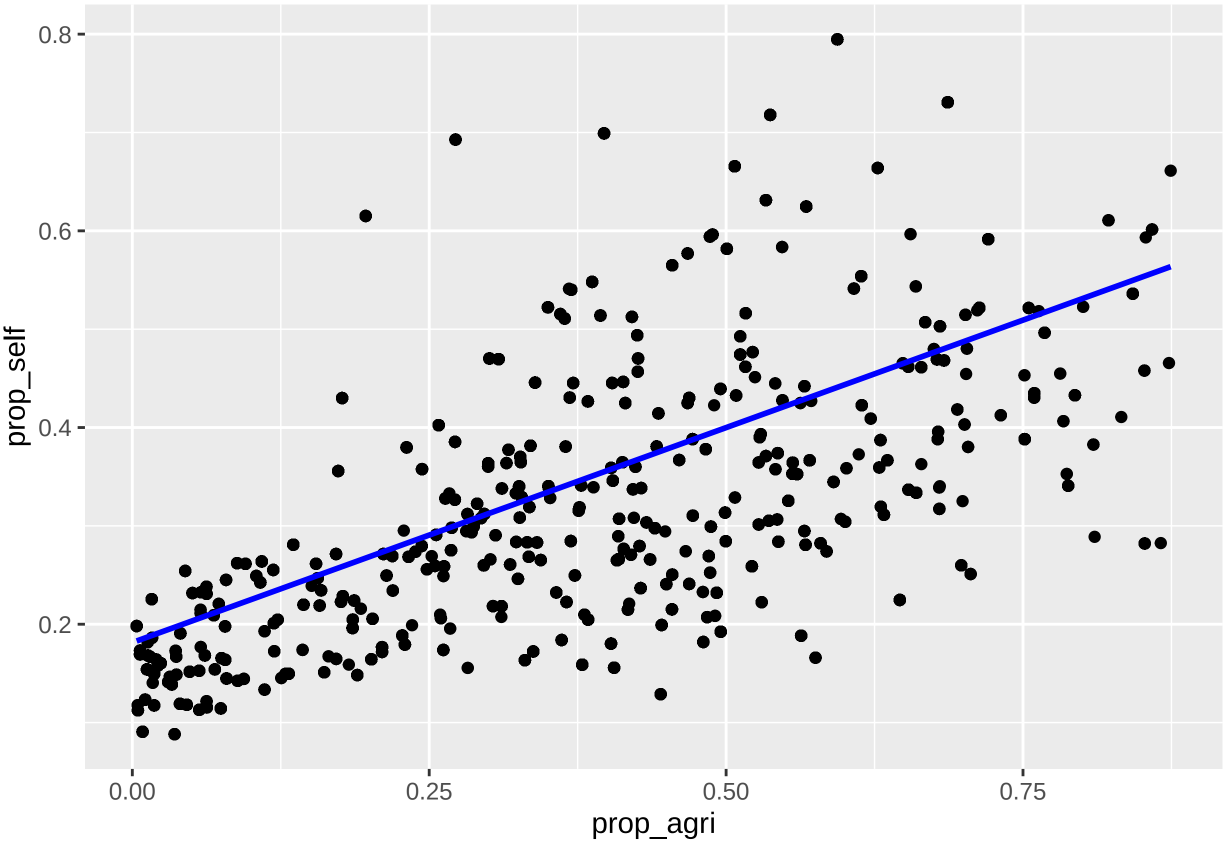
Answer 3a

```r
# Code for 3a
self.agri.mod1 <- lm(prop_self ~ prop_agri, data = migration)
summary(self.agri.mod1)
```

```
##
## 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
```

```
migration %>%
  ggplot(aes(y = prop_self, x = prop_agri)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue")
```

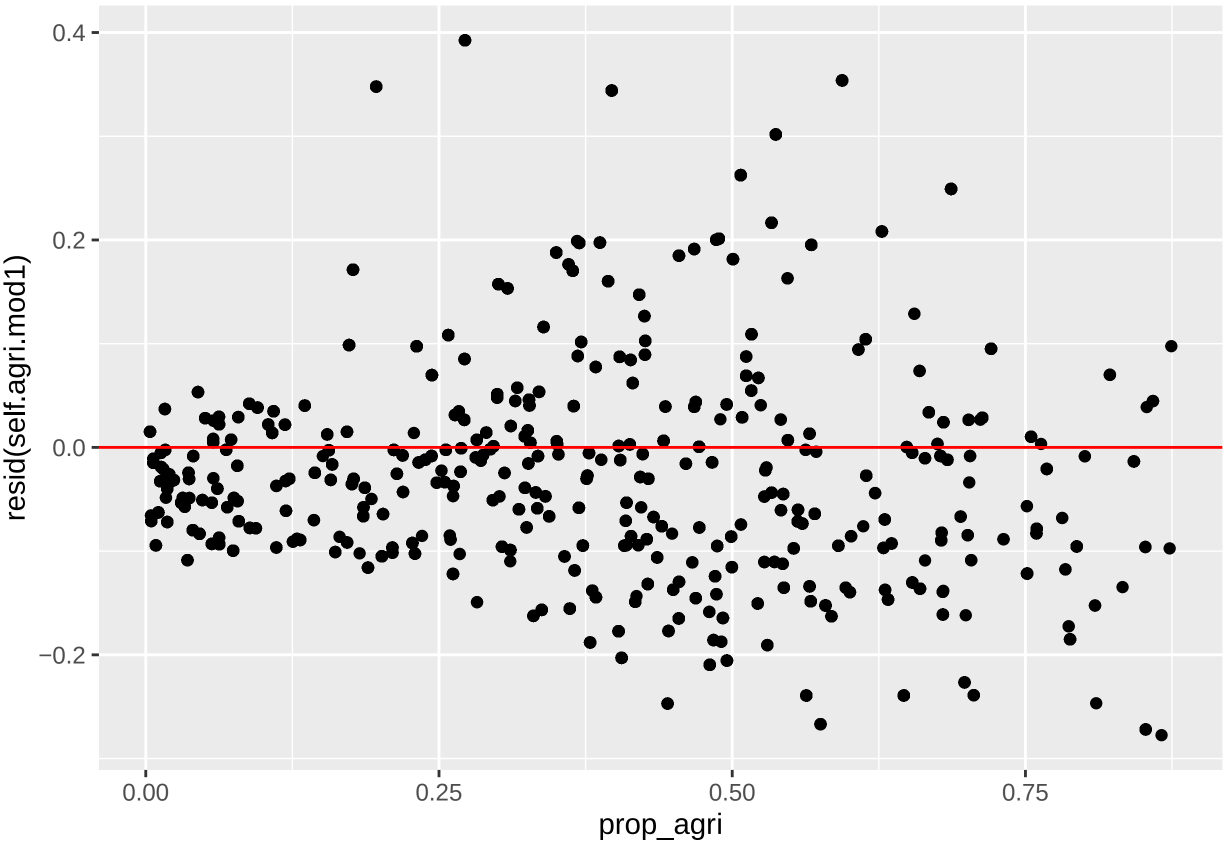
```
## `geom_smooth()` using formula 'y ~ x'
```



### Text Answer 3a

Answer 3b

```
# Code for 3b
migration %>%
  ggplot(aes(x = prop_agri, y = resid(self.agri.mod1))) +
  geom_point() +
  geom_hline(yintercept = 0, color = "red")
```



Text Answer 3c

Scatter plot shows a medium positive association between proportion of community involved in agriculture and the prorpotion who is self-employed. The mean of the residuals appears to be approximately zero in the middle five segments (from 12% to 75%) of the residual plot, which suggest that the linearity assumption holds. In the segment between 0 and 12% proportion of community involved in agriculture and the segment above 75%, the residual plot looks different to the middle segments of the residual plot as more residuals are below the zero line in these segments, suggesting that the mean of the residuals in these segments might not be zero and might need further investigation identifying any other factor impacting the lower and higher end of this residual plot. However, collectively, there is no clear violation of the linearity assumption in this residual plot and we may conclude that linearity assumption holds in this biavariate relationship.

Question 4 [15 pts]

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

4a (5 points)

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

4b (3 points)

Interpret the value of the slope coefficient: Describe what this number tells you in words. Describe the slope on a meaningful scale.

4c (2 points)

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, predict the proportion of self-employed workers there are in this new respondent's community.

4d (3 points)

State and interpret the value of the RMSE and relate it to your answer to 4c (what can you say about the precision of that estimate?).

4e (2 points)

State and interpret the R^2 of the model.

Answer 4

Txxt Answer 4a

$$propselfempl_i = 0.181 + 0.437 * propagri_i + \hat{\epsilon}_i$$

The y-intercept is the expected proportion of the community which is self_employed and is about 18.1%, for the 0% proportion of community which is involved in agriculture. It can be practically meaningful at community level where self-employed people might not be involved in agriculture and may be doing some other work, therefore, in this case, y-intercept is meaningful.

Text Answer 4b

Here, we are comparing the proportions therefore interpretation in percentages will make more sense. Here, the slope indicates that for a 1 unit change in the proportion of population involved in agriculture - where a change of one unit means a 100% change - we expect a 0.437 increase in proportion of community which is self-employed - means 43.7% increase. On a meaningful scale, for every increase of 100 percentage points in the proportion of community involved in agriculture, there is 43.7 percentage points increase the self-employed proportion of community.

Text Answer 4c

```
0.181 + 0.437*(0.2)
```

```
## [1] 0.2684
```

```
predict(self.agri.mod1, newdata = tibble(prop_agri = 0.2))
```

```
##         1
## 0.2687999
```

For an addition of a new respondent to the survey in a community where the proportion of workers involved in agriculture is 0.2 or 20%, we predict that the total proportion of self-employed workers will be approximately 0.27 or 27%. ### Text Answer 4d Residual standard error or RMSE: 0.1238. In this model, RMSE can be interpreted as the average proportion of community which is self-employed is above or below than we would expect based on the proportion of community involved in agriculture. For this model we see that the RMSE is 0.124, meaning that the average proportion of community which is self-employed is 12.4% above or below than we would expect based on the proportion of community involved in agriculture. In relation to to answer in 4c, we can say that for an addition of a new respondent to the survey in a community where the proportion of workers involved in agriculture is 20%, on average, we should expect that the predicted proportion of self-employed will be off (above or below) by about 12.4% and can be 14.6% or 39.4%.

Text Answer 4e

R-squared: 0.2929. 29.3% of the uncertainty in self-employed proportion of community is accounted for by the proportion of community involved in agriculture and the linear relationship between self-employed proportion and agri-involved proportion of community.

## Question 5 [15 pts]

In this problem you will use linear regression to investigate the bivariate 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).

5a (5 points)

Repeat the steps described in Question 3 to determine if this bivariate relationship can be usefully modeled by a linear regression (but now 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 (5 points)

Write out the estimated regression equation and interpret the estimated slope and Y-intercept.

5c (3 points)

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 predict 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 (2 points)

State and interpret the Rˆ2 of the model

Answer 5

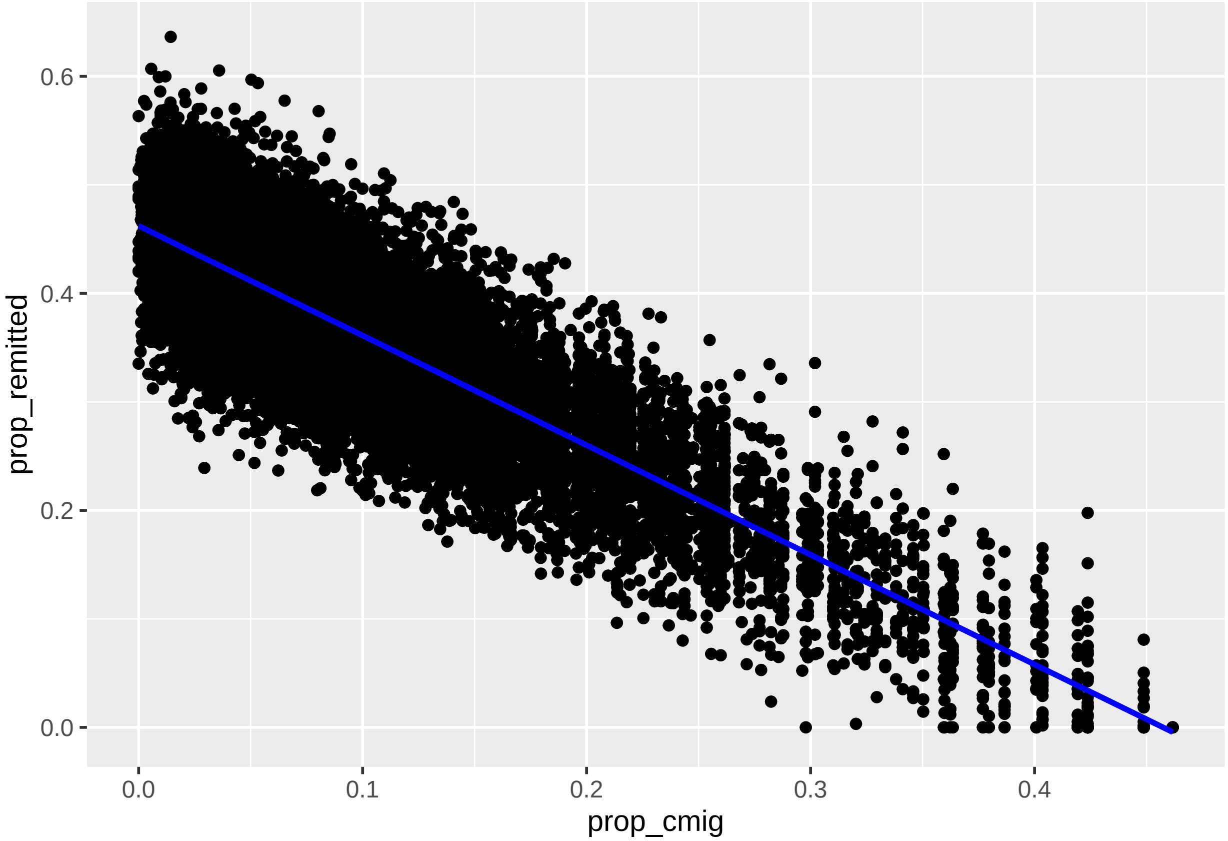
Answer 5a

```
# Code for 5a
cmig.remit.mod1 <- lm(prop_remitted ~ prop_cmig, data = migration)
summary(cmig.remit.mod1)
```

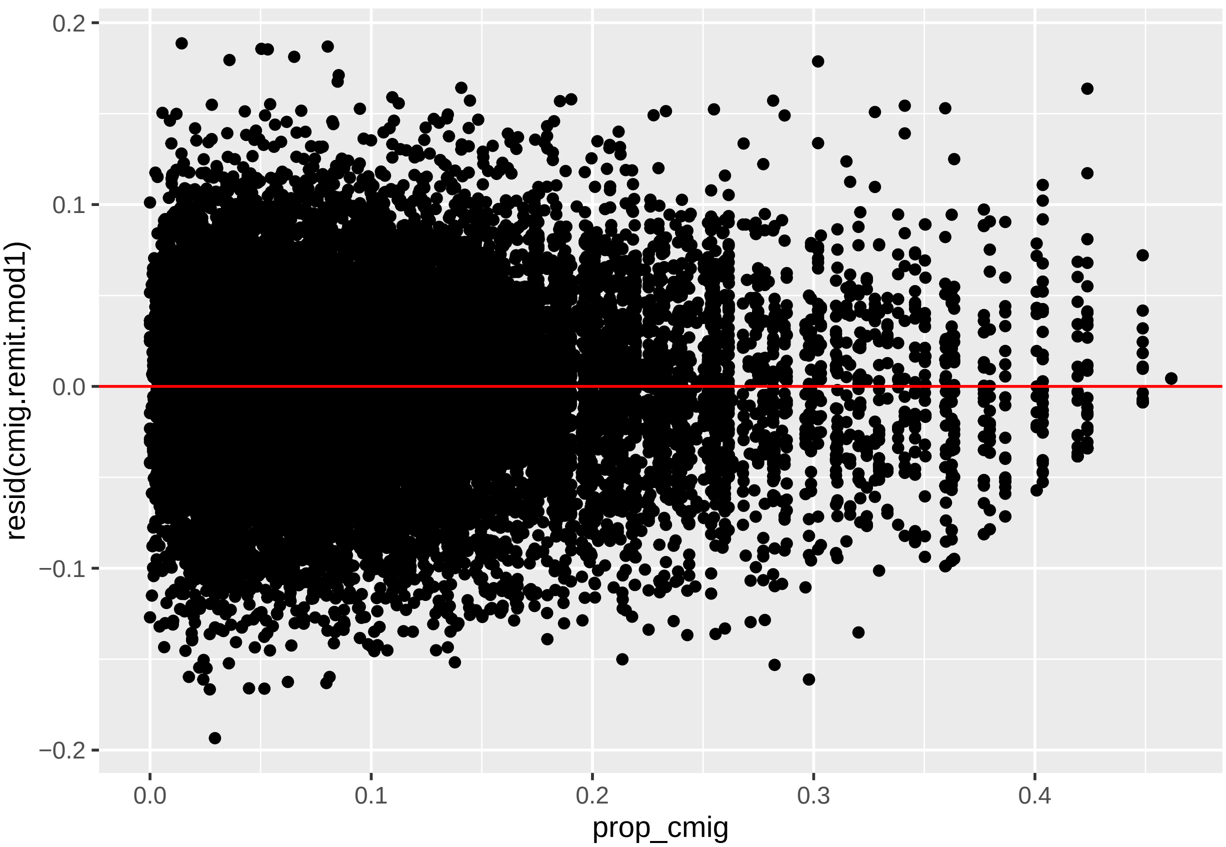
```
##
## Call:
## lm(formula = prop_remitted ~ prop_cmig, data = migration)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.193438 -0.033725  0.000254  0.033827  0.188654
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.4622090  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
```

```
migration %>%
  ggplot(aes(y = prop_remitted, x = prop_cmig)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



```
migration %>%
  ggplot(aes(x = prop_cmig, y = resid(cmig.remit.mod1))) +
  geom_point() +
  geom_hline(yintercept = 0, color = "red")
```



Text Answer 5b

$$propremitted_i = 0.46 - 1.01 * propcmig_i + \hat{\epsilon}_i$$

Here, the estimated slope is -1.01, which means that with a unit increase in the proportion of respondent's community who are also US migrants, a decrease of 1.01 units occur in the proportion of respondent's income remitted to Mexico. On a practical scale, an increase of 10% in proportion of respondent's community who are also US migrants, a decrease of 10.1% in remittance to Mexico occurs - the more migrants, the less remittance to Mexico, also makes sense practically as more of respondent's relatives might be coming to the US and there would be lesser need to remit money to Mexico.

The y-intercept of 46% is the expected proportion of the respondent's income sent to Mexico for the 0% proportion of respondent's community who are also US migrants. It can be practically meaningful at community level where a respondent whose 0% community is US migrant, sends 46% of his income back to Mexico.

Text Answer 5c

```
0.46 - 1.01*(0.15)
```

```
## [1] 0.3085
```

```
predict(cmig.remit.mod1, newdata = tibble(prop_cmig = 0.15))
```

```
##         1
## 0.3106229
```

For an addition of 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 or 15%, we predict that the total proportion of the new respondent's income that is sent back to Mexico in the form of remittances is 0.311 or 31.1%.

Residual standard error or RMSE: 0.04985. In this model, RMSE can be interpreted as the average proportion of respondent's income remitted to Mexico is above or below than we would expect based on the proportion of respondent's community who are also US migrants. For this model we see that the RMSE is 0.05, meaning that the average proportion of respondent's income remitted to Mexico is 5% above or below than we would expect based on the proportion of respondent's community who are also US migrants.

Text Answer 5d

R-squared: 0.719. 72% of the uncertainty in proportion of respondent's income remitted to Mexico is accounted for by the proportion of respondent's community who are also US migrants and the linear relationship between proportion of respondent's income remitted to Mexico and proportion of respondent's community who are also US migrants.