Exploring Relationships in Mexican Migration Project Data

This exercise was created by Dr. Jacqueline Mauro (CMU Stat+Policy PhD graduate) 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
`p_remitted` Pro	portion 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
`p_cmig` Per	cent of respondent's community who are also U.S. migrants
`log_npop`	Logged size of respondent's community.
`p_self` Per	cent of respondent's community who are self-employed
`p_agri` Per	cent of respondent's community involved in agriculture
`p_lessminwage` Per	cent of respondent's community who earn less than the U.S. minimum wage

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

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
```

```
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
# Load data
migration <- read.csv("data3/migration.csv")
# Convert proportions (0 to 1) to percents (0 to 100)
migration$p_remitted <- 100*migration$prop_remitted
migration$p_cmig <- 100*migration$prop_cmig
migration$p_self <- 100*migration$prop_self
migration$p_agri <- 100*migration$prop_agri
migration$p_lessminwage <- 100*migration$prop_lessminwage
```

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 (other summary statistics or features of the distributions of the characteristics in the dataset) it would be most useful to add to your knowledge of the means and why each is important.

Answer 1

Answer 1a

```
mean(migration$year)
```

```
## [1] 1985.832
```

```
mean(migration$age)
```

```
## [1] 24.24353
```

```
mean(migration$p_remitted)
```

```
## [1] 35.61234
```

```
mean(migration$male)
```

```
## [1] 0.7202182
```

```
mean(migration$educ)
```

```
## [1] 6.793536
```

```
summary(migration$year)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1966    1979    1986    1986    1992    2002
```

```
sd(migration$year)
```

```
## [1] 8.647521
```

```r
sd(migration$age)
```

```
## [1] 8.241924
```

```r
summary(migration$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   18.00   22.00   24.24   28.00   65.00
```

```r
summary(migration$p_remitted)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   30.44   36.97   35.61   42.13   63.64
```

```r
summary(migration$educ)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   5.000   6.000   6.794   9.000  25.000
```

```r
mean(migration$p_agri)
```

```
## [1] 37.43833
```

```r
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   p_remitted        p_cmig
##  Min.   :0.003632   Min.   :0.1298   Min.   : 0.00   Min.   : 0.000
##  1st Qu.:0.261930   1st Qu.:0.1319   1st Qu.:30.44   1st Qu.: 4.582
##  Median :0.372650   Median :0.1373   Median :36.97   Median : 8.669
##  Mean   :0.374383   Mean   :0.1386   Mean   :35.61   Mean   :10.498
##  3rd Qu.:0.495250   3rd Qu.:0.1455   3rd Qu.:42.13   3rd Qu.:14.228
##  Max.   :0.874364   Max.   :0.1565   Max.   :63.64   Max.   :46.166
##      p_self           p_agri        p_lessminwage
##  Min.   : 8.821   Min.   : 0.3632   Min.   :12.98
##  1st Qu.:23.424   1st Qu.:26.1930   1st Qu.:13.19
```

```
##  Median :32.665   Median :37.2650   Median :13.73
##  Mean   :34.503   Mean   :37.4383   Mean   :13.86
##  3rd Qu.:43.930   3rd Qu.:49.5250   3rd Qu.:14.55
##  Max.   :79.469   Max.   :87.4364   Max.   :15.65
```

```
summary(migration$p_agri)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3632 26.1930 37.2650 37.4383 49.5250 87.4364
```

```
sd(migration$p_agri)
```

```
## [1] 18.22187
```

```
mean(migration$p_cmig)
```

```
## [1] 10.49755
```

```
sd(migration$p_cmig)
```

```
## [1] 7.889917
```

```
mean(migration$p_self)
```

```
## [1] 34.50283
```

```
mean(migration$p_lessminwage)
```

```
## [1] 13.85603
```

```
mean(migration$log_npop)
```

```
## [1] 8.923845
```

Text Answer 1a

An individual migrant is of age 24 years and a 70% chance of being male with 6-7 years of education and the percentage of income sent as remittences is 35%. the year that they first travelled to the united states would be around 1985. at the community level: the respondant is from a town that has 10.49% of its population who are US immigrants. 37.43% of migrants from a respondants community are involved with agriculture. 34.50% of migrants from a respondant's community would be self employed. 13.85 % of migrants would be earning less minimum wages than in the US. ### Text Answer 1b the combination of means for the individual level and the community level are similar to the median of the groups and hence can be useful to describe an individual immigrant.however, it would be useful to have the standard deviations of the variables along with the mean. It would also be helpful to have the median and the IQR for data that can be skewed as the median and the IQR are not heavily influenced by the outliers. Further the IQR can determine the middle 50 percent of the data.it is also useful to view the distribution as a histogram or boxplot for the skew and to determine how many outliers there could be.

Question 2 [11 pts]

2a (8 points)

Create scatterplots to investigate the bivariate relationship between `p_self` and `p_agri`, as well as the bivariate relationship between `p_self` and `log_npop`. In both figures put `p_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 a higher percent of people are self-employed informative about (predictive of) these two other aspects of their area?

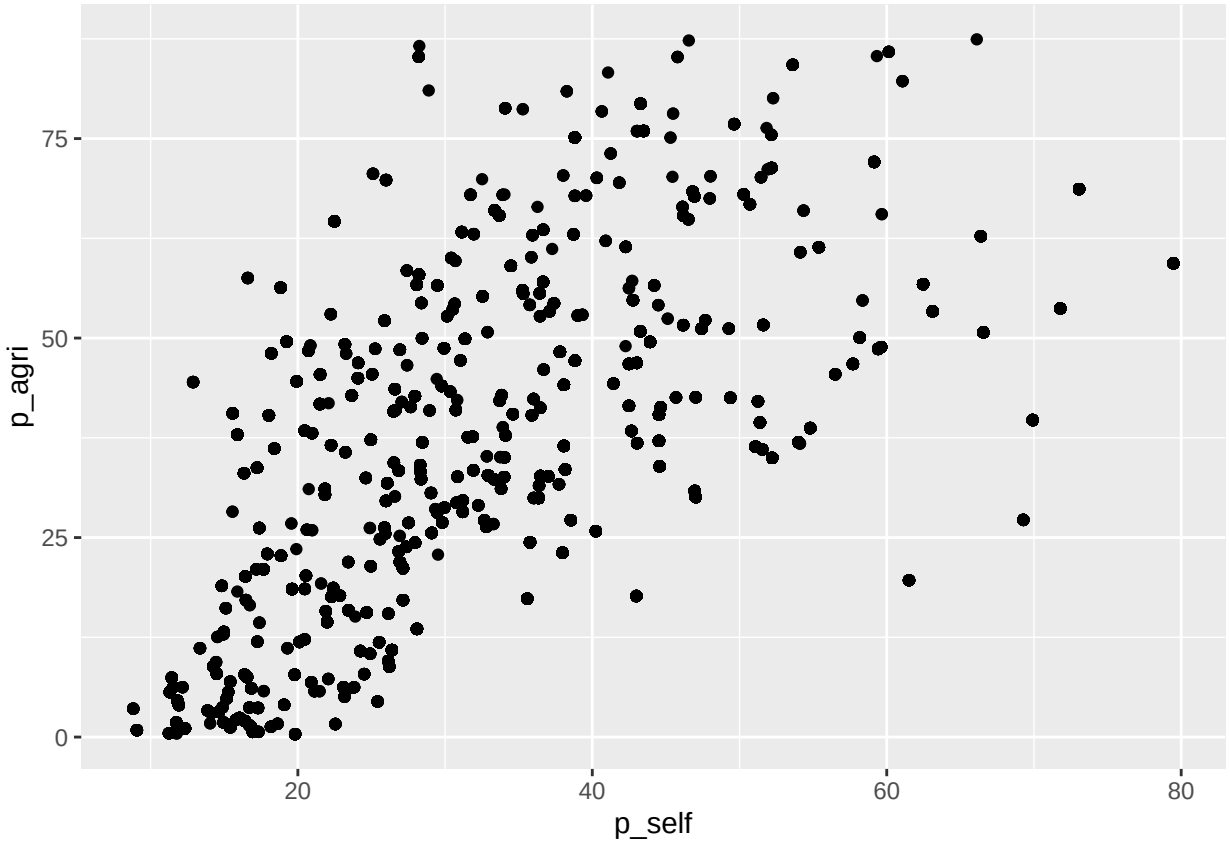
2b (3 points)

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

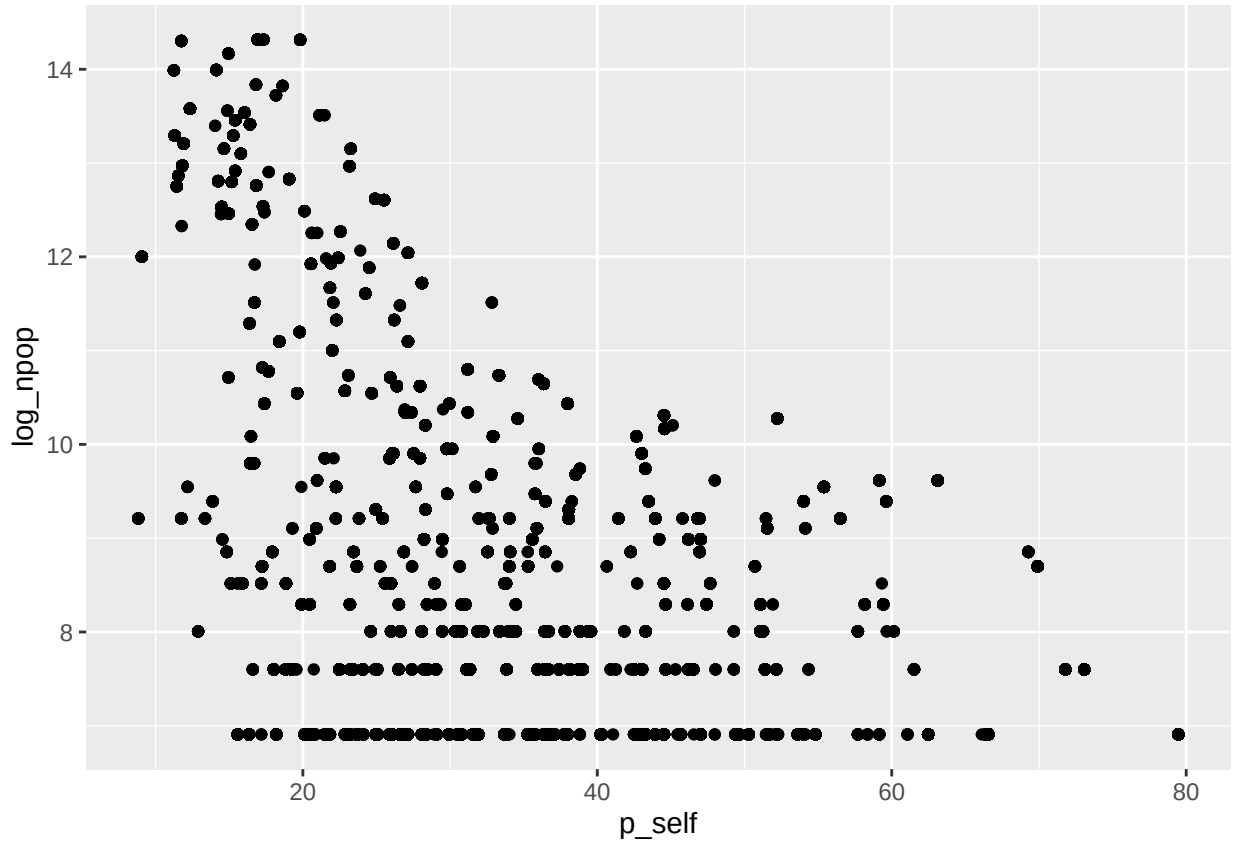
Answer 2

Answer 2a

```
# Code for 2a
migration_com <- data.frame(p_self = migration$p_self, p_agri = migration$p_agri, migration$p_cmig, migr
migration %>%
ggplot(aes(y = p_agri, x = p_self)) +
geom_point()
```



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



Text Answer 2a there is a weak positive correlation observed between the p_self and the p_agri variables. individuals from towns that have more proportion of self employed tend to also do agriculture. There is a slightly negative correlation observed between the p_self and the log_npop. individuals that seem to come from places where there is lower percentage of self employment between 10-30 p_self values are also observed to have higher community size.There seems to be individuals of community size of 7 where individuals have no correlation with the p_self values. there is one outlier with a 79% migrant self - employment percentage that are from a community size of 7.individuals with higher percentage in self employment ie between 40- 79 have lower community sizes.if an individual is from a region with higher percentage of self employment rates they might be invested in agriculture and have lower sizes of communities, ### Answer 2b

```
cor(migration$p_agri,migration$p_self)
```

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

```
migration_com %>%
  select_if(is.numeric) %>%
  cor()
```

```
##                            p_self      p_agri migration.p_cmig
## p_self                  1.0000000  0.54115979       0.29811153
```

```
## p_agri                    0.5411598  1.00000000       0.03898869
## migration.p_cmig          0.2981115  0.03898869       1.00000000
## migration.p_lessminwage -0.1079667  0.37386371      -0.28170913
## migration.log_npop      -0.4319743 -0.65214371      -0.27400955
##                         migration.p_lessminwage migration.log_npop
## p_self                               -0.10796669        -0.43197430
## p_agri                                0.37386371        -0.65214371
## migration.p_cmig                     -0.28170913        -0.27400955
## migration.p_lessminwage               1.00000000        -0.05677052
## migration.log_npop                   -0.05677052         1.00000000
```

Text Answer 2b

the strongest is between log_npop and the p_agri.Therefore there is a strong correlation between the size of the community of the correspondant and the percentage of that correspondant's town that pursues agriculture.

Question 3 [8 pts]

3a (3 points)

Check if the relationship between the percent of people in a migrant's community who are self-employed and the percent 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 percent of self-employed people in the community (this is the outcome or response variable) on the percent 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

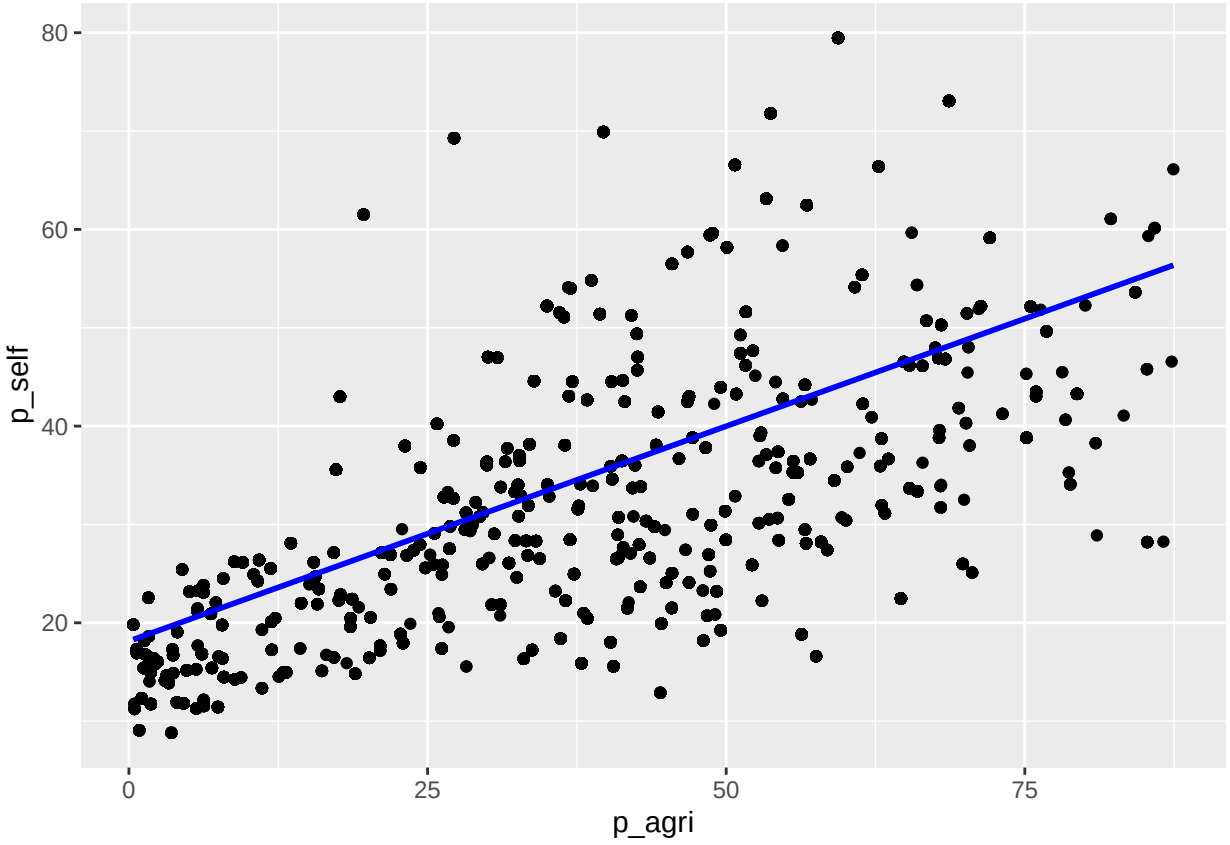
Answer 3a

```
migration_mod <- lm(p_self ~ p_agri, data = migration)
summary(migration_mod)
```

```
##
## Call:
## lm(formula = p_self ~ p_agri, data = migration)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.745  -9.467  -0.836   6.906  39.246
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 18.137362   0.216619   83.73   <2e-16 ***
## p_agri       0.437132   0.005203   84.02   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.38 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
```

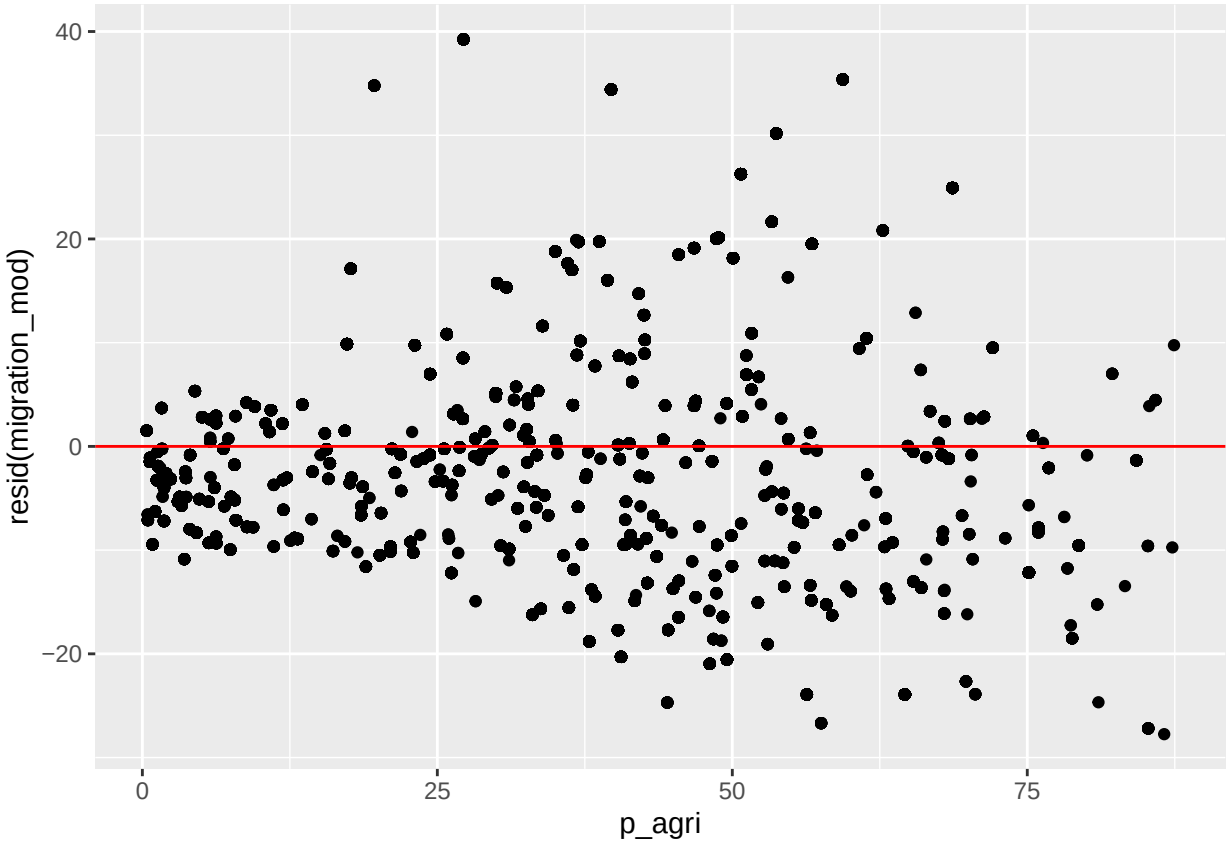
```r
migration %>%
ggplot(aes(y = p_self, x = p_agri)) +
geom_point() +
geom_smooth(method = "lm", se = FALSE, color = "blue")
```

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



Answer 3b

```r
migration %>%
  ggplot(aes(x = p_agri, y = resid(migration_mod))) +
  geom_point() +
  geom_hline(yintercept = 0, color = "red")
```




```
predict(migration_mod, newdata = tibble(p_agri = 20))
```

```
##        1
## 26.87999
```

Text Answer 3c

The mean of the residuals appears to not be zero in each segment of the residual plot as the percent of community involed in agriculture(p_agri) values increase, suggesting that the linearity assumption does not hold. The residual plot between 0-25 seems to have a mean that is not zero. there are more residuals below the zero line than above, and the residuals above the zero line are few in number between 0-25.suggesting that the mean is not likely to be near zero. similarly the last segment has unequal distribution above and below and the mean is not zero.

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 percent of workers involved in agriculture is 20 percent. Using the linear regression results, predict the percent 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

Text Answer 4a

(P-self)Yi= 18.13 - 0.43Xi(p- agri) + Ei when the value of Xi= 0 which is the predictor and the percent of the community who pursue agriculture is zero then we expect that the outcome variable which is the P-self or the percentage of migrants involved in self employment in the region to be 18.13 %. No it is not meaningful as there are no regions that have a 0 percentage for their p_agri or percentage of migrants community in agriculture. there are a few which have a close to zero agriculture percentage we can expect the outcome of self employment percent in those regions to maybe take up values close to 18.

Text Answer 4b

the slope indicates that for a 1 percent increase in the percentage of the corresponds community engaged in agriculture (1 %), we expect the communities to have a 0.43 percent point increase in the correspondants community who are self employed (0.43%). This means a 10 percent point increase in the percentage of community engaged in agriculture is associated with a 4.3 percentage point increase in the percentage of migrants that are self employed in that region. ### Text Answer 4c

the predicted value for a new data point: p_agri= 20 is 26.87%

Text Answer 4d

The RMSE is the average distance the points in the dataset are from the regression line. In this application, we can interpret it as the average percent more or less that the percentage of migrant community that are self- employed in that region than we would expect based on their percentage in pursuing agriculture. For this model, we see that the RMSE is 12.38, meaning that the percentage of correspondants community that are self- employed are on average 12.38 percentage points away from what we would expect based on the p_agri values(percetn of migrants involved in agriculture). we can predict that the percentage of community that is self- employed with the percentage of migrants involved in agriculture as 20% will be 26.87%. we should expect this prediction for self-employment percentage in the migrant communities to be off (above or below) on average by about 12.38 percent points. Half the time, the true value for the self- employment percent for communities will be no more than 9 percent points below and 6 percent points above the prediction we make with this regression model. The worst of our predictions could be as far as 27 % points below the true value and 39.24% above the true value.

Text Answer 4e

The R2 is 0.29, meaning that 29 percent of the variability in percentage of self-employment in correspondants community can be explained by the predictor p_agri(percent of community involved in agriculture) and the linear relationship between p_agri variable(the percentage of community in agriculture) and the p_self(percentage of community that are self-employment). 42% of this estimate for the percent of migrants that are self employed can be explained by the p-agri values (percent of community involved in agriculture) and the linear relationship between percent of community involved in agriculture(p_agri) and the percent of community that are self -employed(p_self).

Question 5 [15 pts]

In this problem you will use linear regression to investigate the bivariate relationship between the percent 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 percent 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 (for 5a: Y = percent of a migrant's income that is sent back to Mexico in the form of remittances and X = the percent 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 percent of people in a migrant's community who are also migrants is 15 percent. Using the linear regression results, what do you predict the percent 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 R^2 of the model

Answer 5

Answer 5a

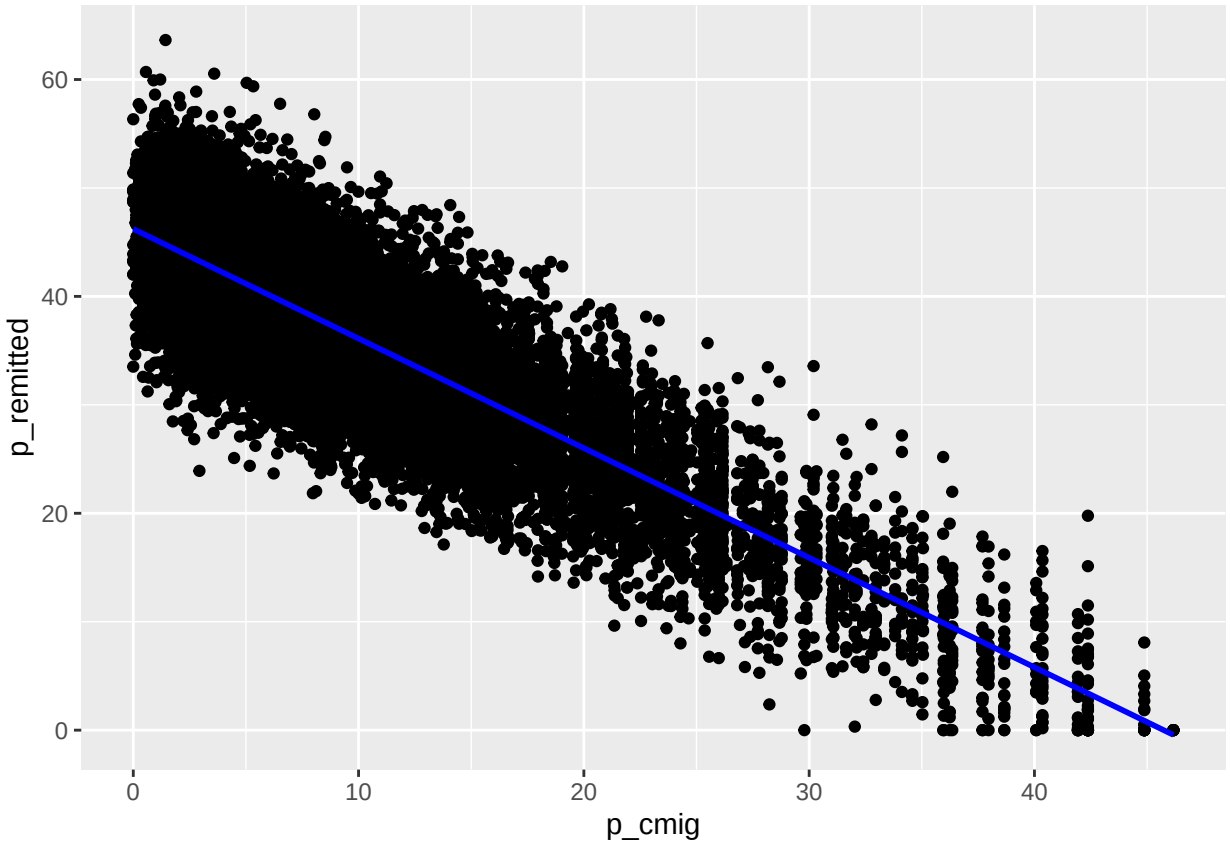
```
migration_mod_5 <- lm(p_remitted ~ p_cmig, data = migration)
summary(migration_mod_5)
```

```
##
## Call:
## lm(formula = p_remitted ~ p_cmig, data = migration)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.3438  -3.3725   0.0254   3.3827  18.8654
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 46.220899   0.063549   727.3   <2e-16 ***
## p_cmig      -1.010574   0.004839  -208.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

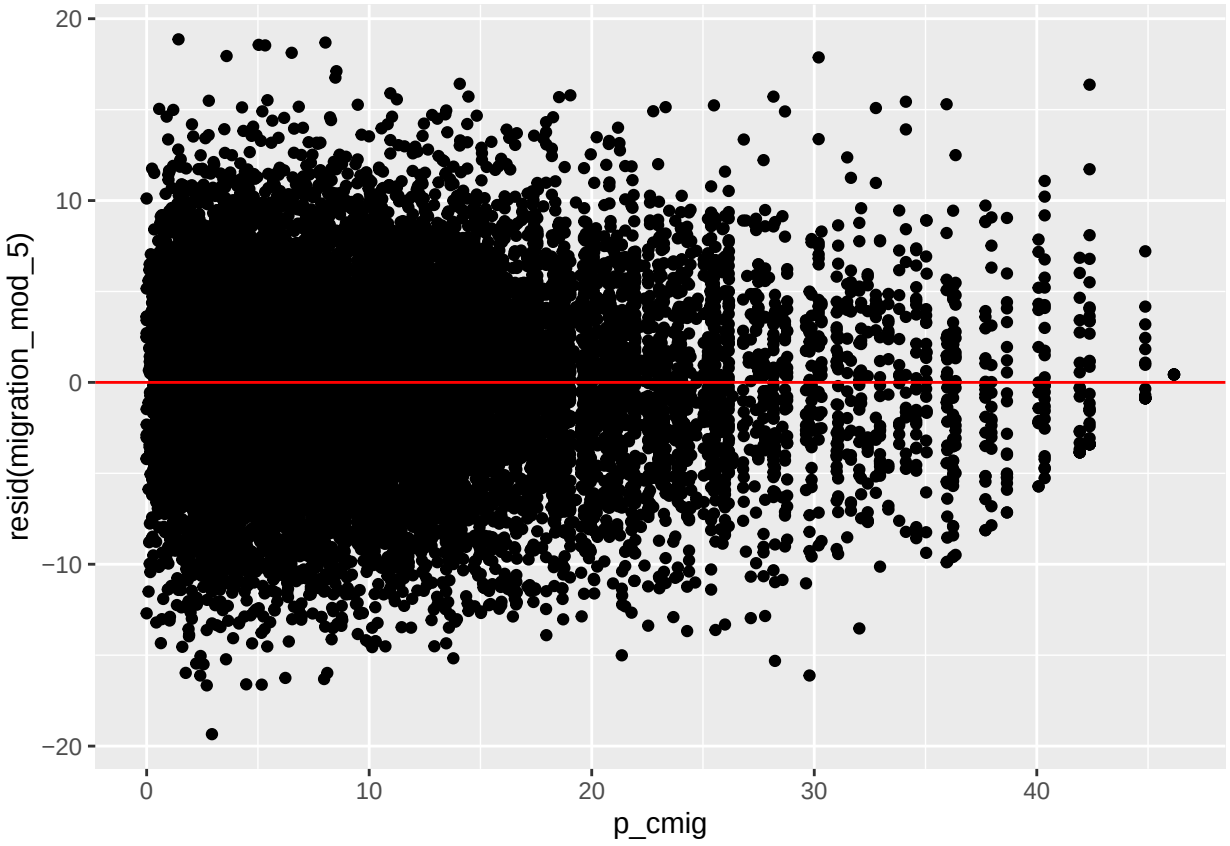
```
##
## Residual standard error: 4.985 on 17047 degrees of freedom
## Multiple R-squared:  0.719,  Adjusted R-squared:  0.7189
## F-statistic: 4.361e+04 on 1 and 17047 DF,  p-value: < 2.2e-16
```

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

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




```
migration %>%
  ggplot(aes(x = p_cmig, y = resid(migration_mod_5))) +
  geom_point() +
  geom_hline(yintercept = 0, color = "red")
```




```
predict(migration_mod, newdata = tibble(p_agri = 20))
```

```
##        1
## 26.87999
```

```
predict(migration_mod_5, newdata = tibble(p_cmig = 15))
```

```
##        1
## 31.06229
```

Text Answer 5a

yes the (p_remittance) percent of a migrant's income that is sent back to Mexico in the form of remittances and the (p_cmig) the percentage of migrants in the community that are US migrants can be modelled and will be useful as the segments between 0-10, 10-20,20-35, 35-45 are all having a mean of zero with equally distributed points above and below the mean.The mean of the residuals appears to be zero in each segment of the residual plot as the percent of people in migrants community increase,suggeting the linearity assumption holds.

Text Answer 5b

(p_remittance)Yi= 46.22-1.01Xi(p_cmig) + Ei

when the value of Xi= 0 which is the predictor value of the percentage of the community who are US migrants is zero then we expect that the outcome variable which is (p_remitted) percentage of migrants income sent as remittances in the region to be 46.22 %. It is meaningful as there are regions whose migrant community have 0% as their (p_cmig)percentage of community that are us migrants.

the slope indicates that for a 1 percentage point increase in the percentage of the community who are US migrants (1 %), we expect the communities to have a -1.01 percent point decrease in the percent of migrants income sent back to mexico as remittances. This means a 10 percent point increase in the percentage of community that are US migrants is associated with a -10.1 percent point decrease in the percentage of migrants income sent as remittance.

Text Answer 5c

predict(migration_mod_5, newdata = tibble(p_cmig = 15)) the new remittance for a data point of 15 percent of community who are also US migrants is 31.06. RMSE: For this model, we see that the RMSE is 4.98, meaning that the (p_remittted)percentage of migrants income sent as remittance is on average 4.98% away from what we would expect based on their (p_cmig) percent of community who are also US migrants . we predict that the percentage of migrants income sent as remittance with the c_mig(percentage of community who are also US migrants) 15% will be about 31.06%. we should expect this prediction of 31.08% to be off (above or below) by about 4.98 percent points this is the average prediction error. Half the time, the true value for the percent of migrant income sent as remittance will be no more than 3.37 percent points below and 3.38 percent points above the prediction we make with this regression model. The worst of our predictions could be as far as 19% points below the true value and 18.86% (percent points) above the true value.

Text Answer 5d

The R2 is 0.71, meaning that 71.9% of the variability in percentage of income for migrants sent as remittance can be explained by the percent of the migrants that are us migrants and the linear relationship between the two variables.(p_cmig and p_remitted)