Lab 5

October 8, 2020

30 years ago, did the Mafia give support to the Christian Democrat political party in Sicily and were they rewarded with building contracts?

Name	Description
comune	Geographic area
year	Year of the election
DC_VV	Share of votes won by the Christian Democrats
bui_labf	Proportion of labor force in construction
Mafia1987	Mafia town in 1987
young_p	Share of population under 25
h_wo_water_wc_p	Houses without basic services per capita
illiterate_p	Share of population who are illiterate
laurea_p	Share of population with a university degree
laurefa_p	Share of female population with a university degree

New(er) functions you will need

- summarize_at
- cor
- lm
- summary (as used to obtain information about a model)
- fitted
- residuals (or the abbrevaited version resid)

Question 0

We've got a lot of variables in this dataset, so let's try to make sense of them.

First, load the data.

```
library(tidyverse)
```

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

```
## v ggplot2 3.3.2     v purrr   0.3.4
## 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
```

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

```
mafia <- read_csv('data/mafia.csv')
```

```
## Parsed with column specification:
## cols(
```

```
##   comune = col_character(),
##   year = col_double(),
##   DC_VV = col_double(),
##   bui_labf = col_double(),
##   laurea_p = col_double(),
##   laureaf_p = col_double(),
##   h_wo_water_wc_p = col_double(),
##   illiterate_p = col_double(),
##   young_p = col_double(),
##   Mafia1987 = col_double()
## )
```

Get the means of some of the variables you think are important.

```
mafia %>%
  summarize_at(vars('Mafia1987', 'bui_labf', 'laurea_p'), 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.
```

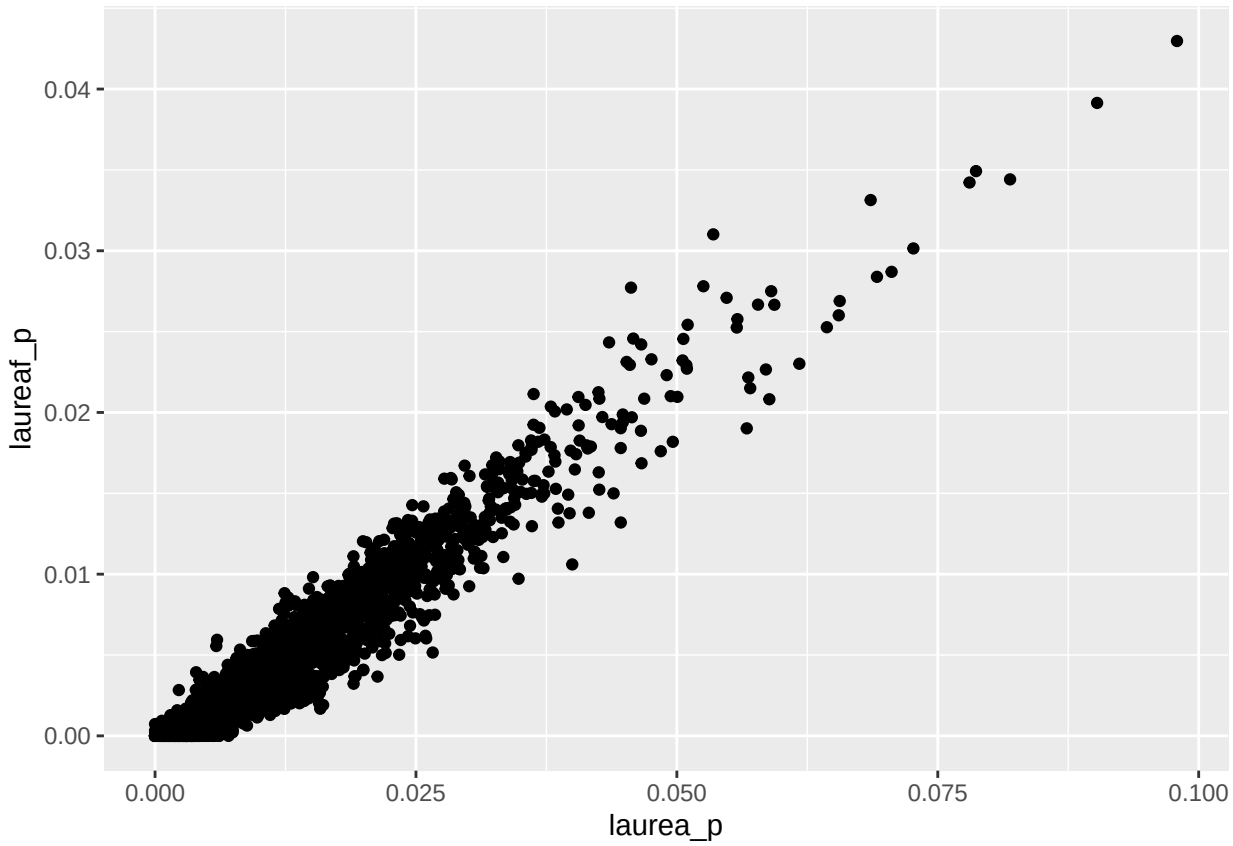
```
## # A tibble: 3 x 2
##   variable   value
##   <chr>      <dbl>
## 1 Mafia1987 0.211
## 2 bui_labf  0.123
## 3 laurea_p  0.0104
```

So 21%of towns were considered under Mafia rule, on average 12% of workers were in construction, and on average only 1% of the population had a university degree in Sicily during this time period.

Check a few relationships between continuous variables using scatterplots. For example, look at the relationship between the share of the everyone in the population of a *comune* who have a university degree and the share of women in a comune with a university degree.

Is this a strong relationship? Does it look linear? What statistic could also help answer those questions?

```
mafia %>%
  ggplot(aes(x = laurea_p, y = laureaf_p)) +
  geom_point()
```




```
cor(mafia$laurea_p, mafia$laureaf_p)
```

```
## [1] 0.9675093
```

Yes, the relationship between proportion of women with a university degree and proportion of everyone with a universtiy degree do seem to have a linear relationship. The linear correlation is about 0.97, this is a very strong linear relationship.

Let's do a quick regression just to remind ourselves what that looks like. We'll regress the share of the population working in construction `bui_labf` (outcome variable) against the share who have college degrees `laurea_p` (predictor variable). What sign do you expect the coefficient to be?

```
reg.prelim <- lm(data = mafia, bui_labf ~ laurea_p)
summary(reg.prelim)
```

```
##
## Call:
## lm(formula = bui_labf ~ laurea_p, data = mafia)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.17531 -0.04895 -0.00602  0.03986  0.32085
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.109669   0.001577   69.56   <2e-16 ***
## laurea_p    1.296322   0.114123   11.36   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.06929 on 4438 degrees of freedom
## Multiple R-squared:  0.02825,    Adjusted R-squared:  0.02803
## F-statistic:   129 on 1 and 4438 DF,  p-value: < 2.2e-16
```

Let's examine what a regression object looks like. Remember that the `str` function we used earlier tells us the structure of any object in R. Notice that when you tell R to print out `reg.prelim`, it only prints a few things. In reality, the object contains a lot more information.

```
reg.prelim
str(reg.prelim)
```

We can extract things from this object.

```
residuals <- reg.prelim$residuals  # or residuals <- resid(reg.prelim)
summary(residuals)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -0.175310 -0.048946 -0.006018  0.000000  0.039862  0.320855
```

Question 1

Describe the distribution of the residuals from Question Zero. What is the RMSE? Plot the residuals against the predictor variable (laurea_p). Does this plots support the idea that the relationship between these two variables is linear? Why or why not?

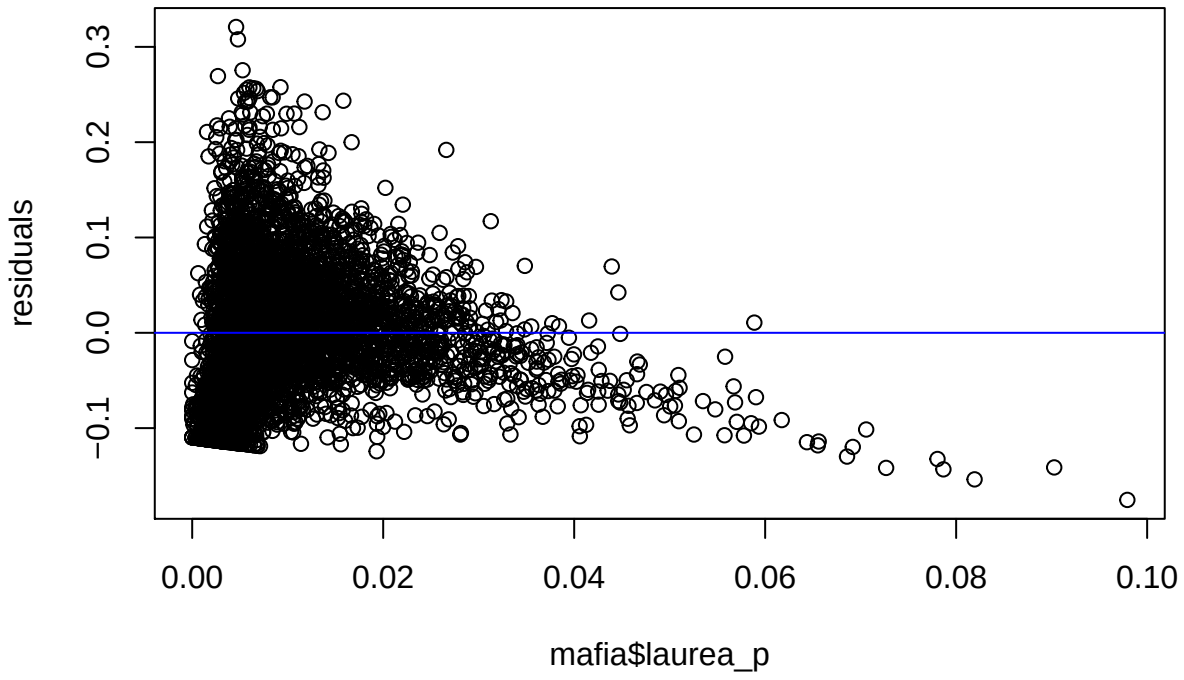
```
summary(residuals)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -0.175310 -0.048946 -0.006018  0.000000  0.039862  0.320855
```

```
sd(residuals)
```

```
## [1] 0.06927768
```

```
plot(mafia$laurea_p, residuals)
abline(h=0, col = "blue")
```



Looked at all together, the residuals are approximately symmetric around zero - the median is very close to the mean. Half the comunes have between 4 percentage points more and 5 percentage points fewer people working in contruction than we expect based on the comune's share of people with a university degree. The RMSE is 0.069 so on average, the proportion of workers in construction in each comune is 7 percentage points away from the estimated regression line. The comune furthest below the line (minimum residual) is 17 percentage points below and the comune farthest above the line (maximum residual) is 32 percentage points above.

From the residual plot we see that the mean of the residuals appears to decrease and fall below zero for comunes whose proportion of those with a university degree is greater than 5%. This violates the linearity assumption and therefore linear regression is not an appropriate or useful model. We therefore should not interpret the slope and intercept coefficients.

Question 2

Check the correlations between a few variables that describe the economic conditions of the *comune*: the share with university degrees, the share without basic services, the share who are illiterate and the share who are under 25. Which seem strongly linearly related, which less so?

Answer 2

```
mafia %>%
  select(laurea_p, laureaf_p, h_wo_water_wc_p, illiterate_p, young_p) %>%
  cor()
```

```
##                   laurea_p  laureaf_p h_wo_water_wc_p illiterate_p    young_p
## laurea_p         1.0000000  0.9675093      -0.3364706   -0.5911373 -0.4425135
## laureaf_p        0.9675093  1.0000000      -0.3274780   -0.6151210 -0.4783976
## h_wo_water_wc_p -0.3364706 -0.3274780       1.0000000    0.6314579  0.3717598
## illiterate_p    -0.5911373 -0.6151210       0.6314579    1.0000000  0.6195809
## young_p         -0.4425135 -0.4783976       0.3717598    0.6195809  1.0000000
```

These seem generally highly correlated, especially the share of women with university degrees and the share of the total population with university degrees, which are strongly positively correlated at 0.9675093. As we'd expect, each of these measures of high educational attainment is in turn negatively correlated with having no basic services and the share of the population that is illiterate. The negative relationship with illiteracy rate is especially strong at roughly -0.6, which makes intuitive sense. There is also a negative linear correlation with the share of the population that is young, also intuitive given you must have reached a certain age to have a university degree.

The measures of poverty are all positively correlated with one another as well. Lack of services and illiteracy are especially strongly correlated at 0.6314579.

From this we can see that in these communities, the markers of poverty all seem to go hand in hand. Some of the lower linear correlations may mask strong non-linear relationships between variables.

Question 3

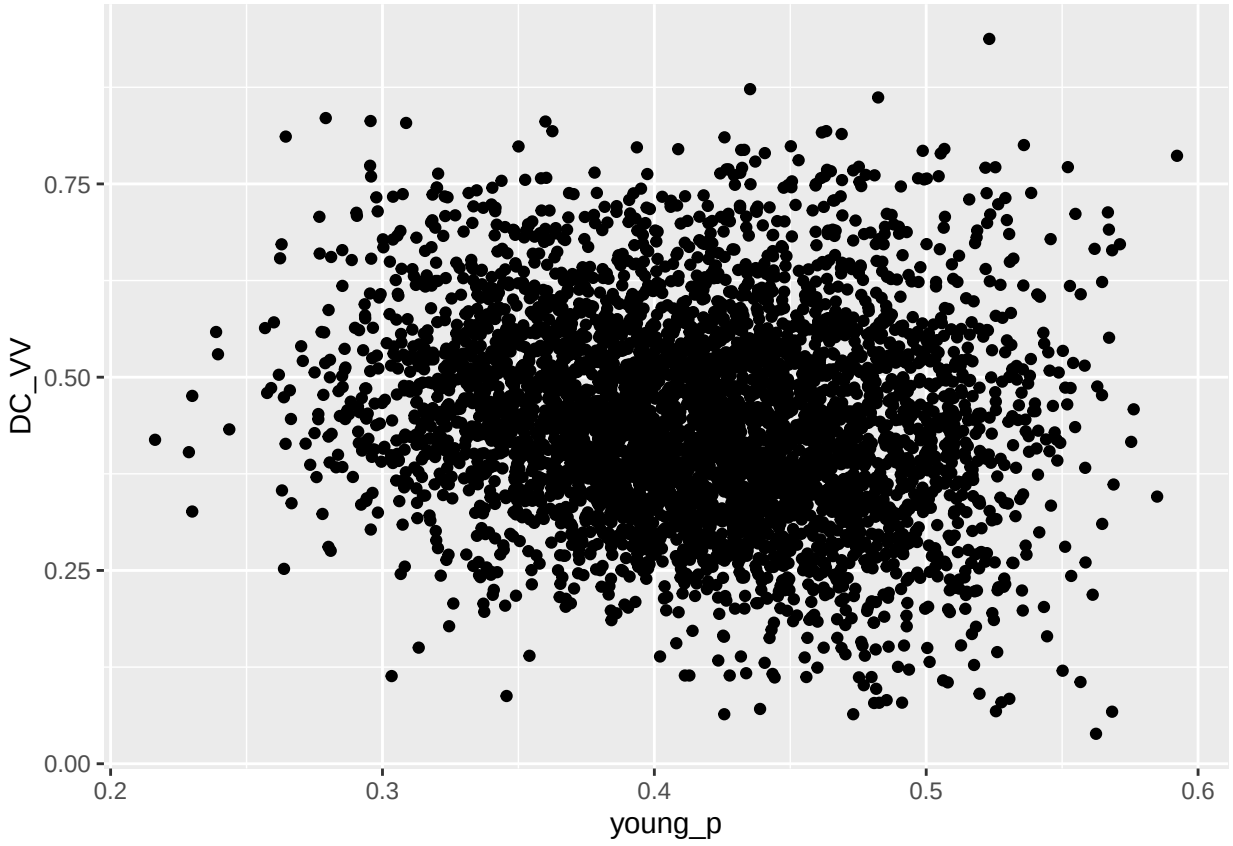
Check if the share of votes for the Christian democratic party increases or decreases with the percentage of young voters by a scatter plot. What do you notice? What do you expect the linear regression to look like?

Next, examine the same question by running a linear regression of the vote shares (outcome variable) against percentage of young voters (predictor variable).

Describe the distribution of the residuals and interpret the RMSE. Plot the residuals against the predictor variable. Do these plots support the idea that the relationship between these variables are linear? If so interpret the coefficients on the regression output.

Answer 3

```
mafia %>%
  ggplot(aes(x = young_p, y = DC_VV)) +
  geom_point()
```



The scatterplot shows very little obvious trend, though possibly slightly negative. We expect the linear regression to have only a small negative slope coefficient, a low coefficient of variation, and a large RMSE.

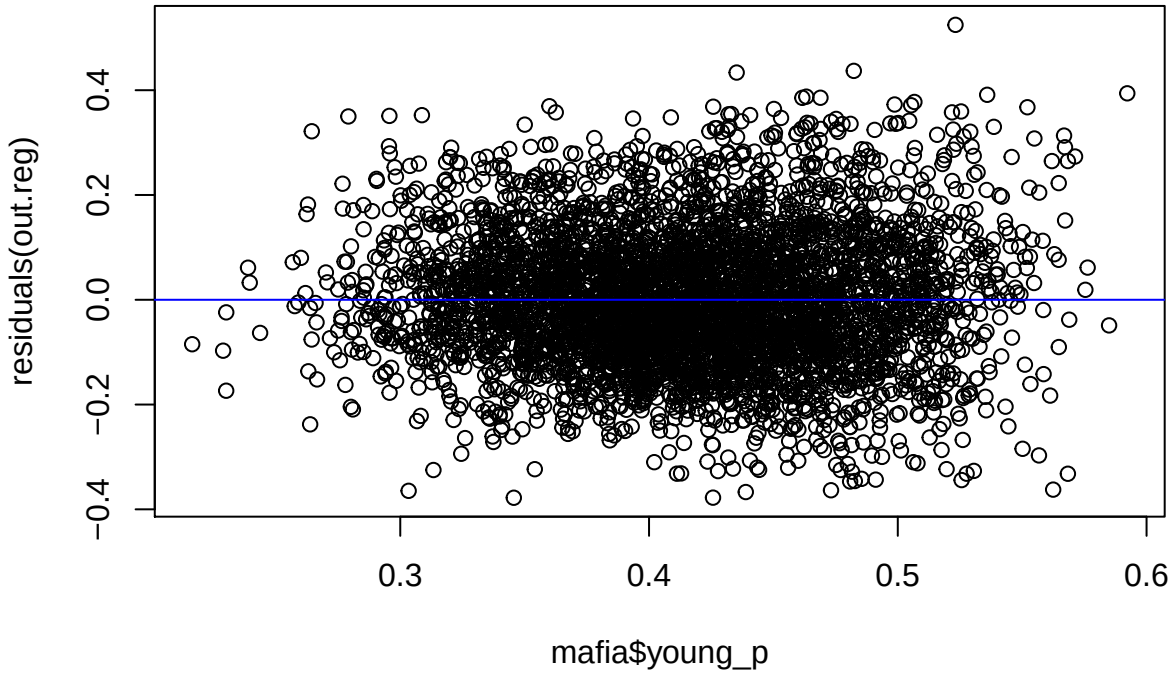
```
out.reg <- lm(DC_VV~young_p, data = mafia)
summary(out.reg)
```

```
##
## Call:
## lm(formula = DC_VV ~ young_p, data = mafia)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.37788 -0.08687 -0.00928  0.07923  0.52475
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.56811    0.01342   42.33   <2e-16 ***
## young_p     -0.29685    0.03199   -9.28   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1275 on 4438 degrees of freedom
```

```
## Multiple R-squared:  0.01904,    Adjusted R-squared:  0.01881
## F-statistic: 86.12 on 1 and 4438 DF,  p-value: < 2.2e-16
```

The RMSE is about 13 percentage points telling us that the share of votes for the Christian democratic party was on avearge 13 percentage points away from what we expected it to be based on its linear relationship with the proportion of the comunes population that was young.Over all comunes, the distribution of residuals appears to be symmetric, the median is close to zero, and half the comunes are within 7-8 percentage points of the regression line. The comune furthest below the line (minimum residual) is 37 percentage points below and the comune farthest above the line (maximum residual) is 52 percentage points above. .

```
plot(mafia$young_p, residuals(out.reg))
abline(h = 0, col = "blue")
```



The residual plot indicates that the mean of the residuals is close to zero for all values of the predictor variable, supporting the idea that linearity assumption holds. We can therefore interpret the model coefficients.

```
summary(out.reg)
```

```
##
## Call:
## lm(formula = DC_VV ~ young_p, data = mafia)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.37788 -0.08687 -0.00928  0.07923  0.52475
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.56811    0.01342   42.33   <2e-16 ***
## young_p     -0.29685    0.03199   -9.28   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1275 on 4438 degrees of freedom
```

```
## Multiple R-squared:  0.01904,    Adjusted R-squared:  0.01881
## F-statistic: 86.12 on 1 and 4438 DF,  p-value: < 2.2e-16
```

The slope coefficient on `young_p` is negative. Interpreting the magnitude of the slope we see that as the percentage of young people increases by 10 percentage points, the Christian Democrats' vote share decreases on average by 3 percentage points.If we increase the proportion of young people from 0 to 1, we would expect the vote share to decrease by 0.298, almost 30 percentage points. The y-intercept indicates that in a region with no young people (persons under 25), we would expect the Christian Democrats to receive 0.568 share of the vote. This is not a meaningful value as no comunes have zero young people.

Question 4

Assume there exists a city with `young_p` = 0.5. Use the regression equation from Question 3 to predict the value of the Christian vote share. Based on the residual standard error, what do you expect the likely prediction error to be?

Answer 4

```
predict(out.reg, newdata = tibble(young_p = 0.5))
```

```
##         1
## 0.4196873
```

```
# another way: intercept = coef(out.reg)[1]; beta1 = coef(out.reg)[2]; level = 0.5
# prediction = intercept + beta1*level
```

Based on our estimated regression line we predict the Christian vote share to be about 0.42. Based on the RMSE 0.13, on average (over multiple predictions) we expect our predictions to be about 13 percentage points away from the true value.

Question 5

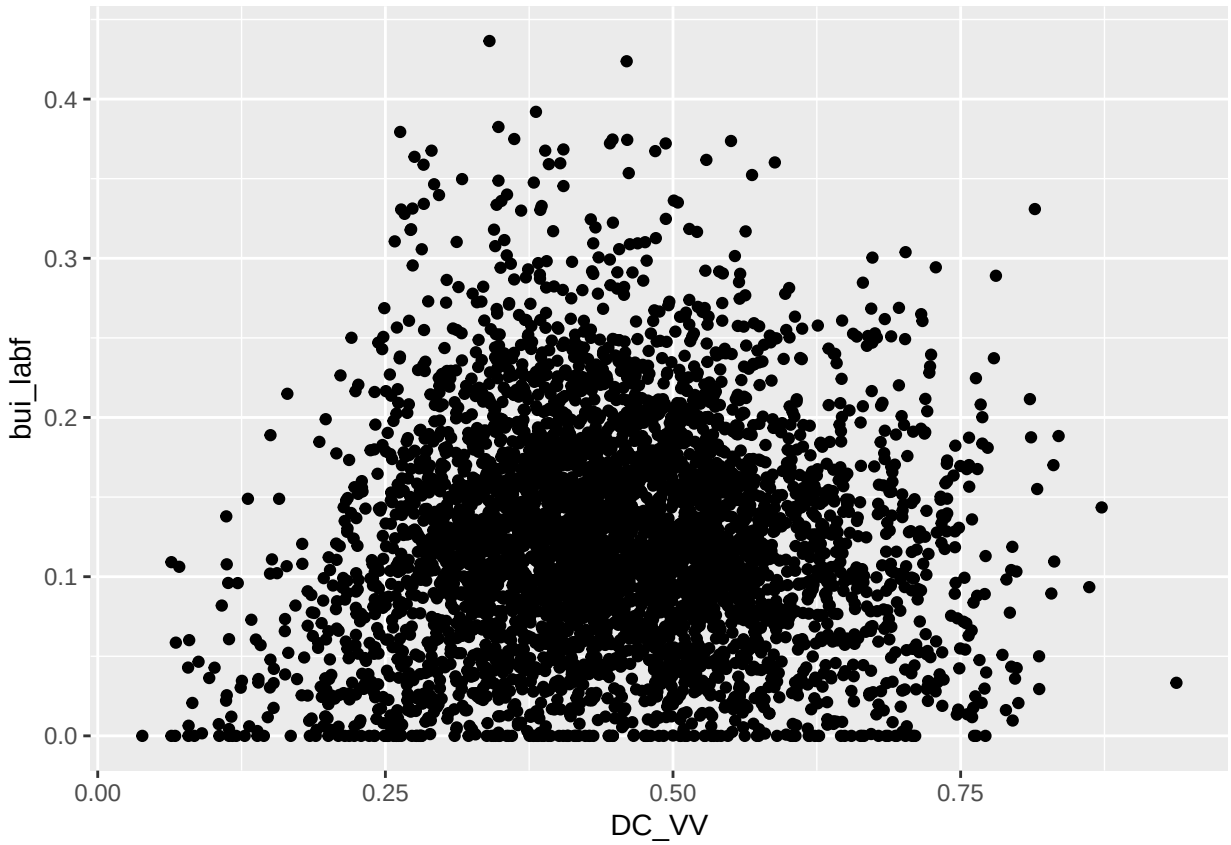
The other question the authors want to examine is whether the Christian Democrats increased funding for construction projects which were generally under the control of the Mafia in Sicily during this time period.

To start examining this, make a scatterplot of `bui_labf` (x-axis) against `DC_VV` (y-axis). Does there appear to be a positive or negative relationship? How strong is this relationship?

Next run a regression of the share of labor in construction (the outcome variable) against the share of votes won by the Christian Democrats (the predictor variable). Briefly comment on the distribution of the residuals, then make a plot of the residuals against the fitted values for this regression. Does this plot support that the true model is linear? If so interpret the model coefficients.

Answer 5

```
mafia %>%
  ggplot(aes(y = bui_labf, x = DC_VV)) +
  geom_point()
```




```
cor(mafia$bui_labf, mafia$DC_VV)
```

```
## [1] 0.04057889
```

There appears to be a slight positive relationship between these variables. It is unclear from the scatterplot whether the relationship is linear. The linear correlation is very close to zero, 0.04.

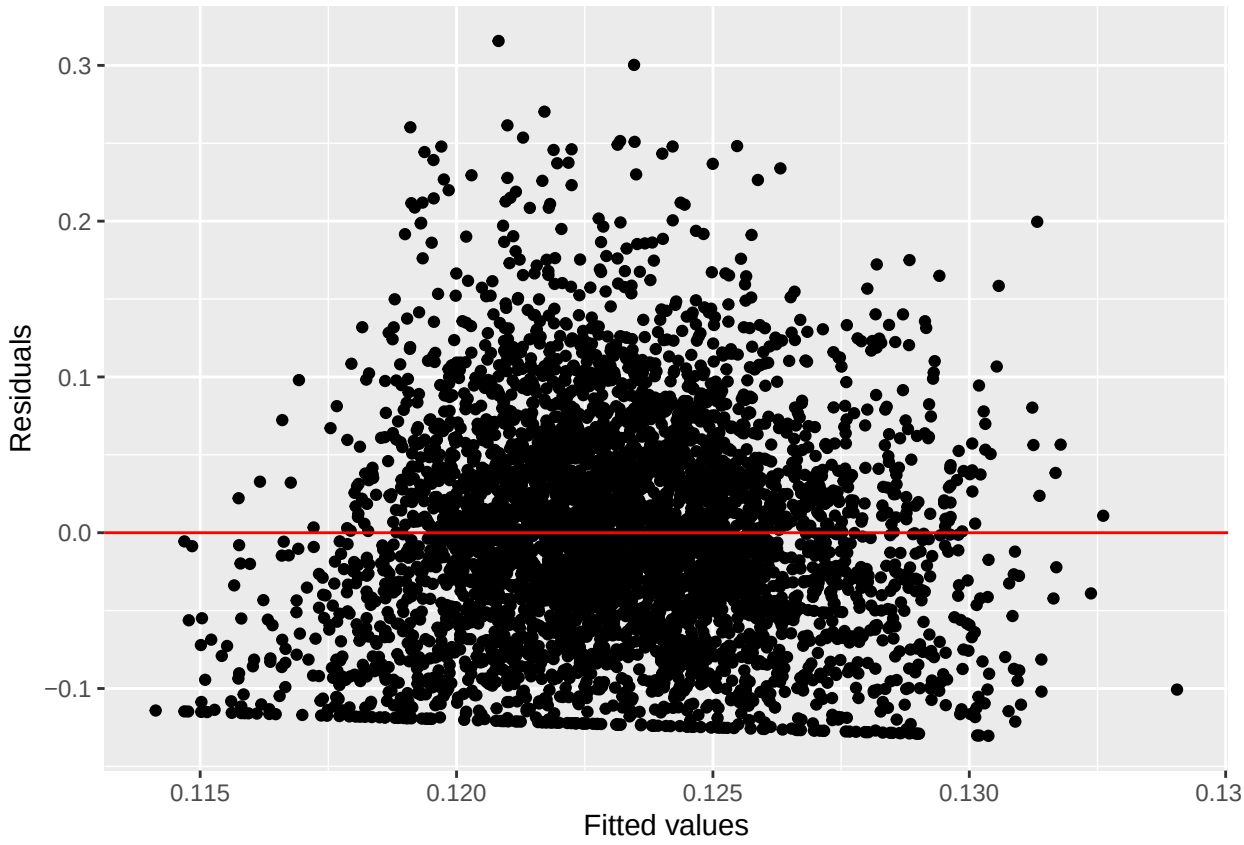
```
const.reg <- lm(data = mafia, bui_labf ~ DC_VV)
summary(const.reg)
```

```
##
## Call:
## lm(formula = bui_labf ~ DC_VV, data = mafia)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.130373 -0.049013 -0.002811  0.041029  0.315665
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.113274   0.003793  29.863  < 2e-16 ***
## DC_VV       0.022162   0.008191   2.706  0.00685 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07023 on 4438 degrees of freedom
## Multiple R-squared:  0.001647,   Adjusted R-squared:  0.001422
## F-statistic:  7.32 on 1 and 4438 DF,  p-value: 0.006846
```

The RMSE is about 7 percentage points telling us that the percent of workers in construction was on avearge 7 percentage points away from what we expected it to be based on its linear relationship with the share of votes for the Christian Democratic party. Over all comunes, the distribution of residuals appears to be symmetric, the median is close to zero, and half the comunes are within 4-5 percentage points of the regression line. The comune furthest below the line (minimum residual) is 13 percentage points below and the comune farthest above the line (maximum residual) is 32 percentage points above.

```
newdat <- tibble(residuals = const.reg$residuals,
                 fitted.values = const.reg$fitted.values)

newdat %>%
  ggplot(aes(x = fitted.values, y = residuals)) +
  geom_point() +
  geom_hline(yintercept = 0, color = 'red') +
  labs(x = "Fitted values", y = "Residuals", main = "Fitted vs. Residuals")
```



The mean of the residuals appear to change depending where you are on the fitted values; for comunes whose predicted outcome is less than 12% or more than 13% the mean of the residuals appears to be negative. This suggests that the linearity assumption is violated (or does not hold). The linear model is not appropriate for this data and it is not useful to interpret the estimated model coefficients