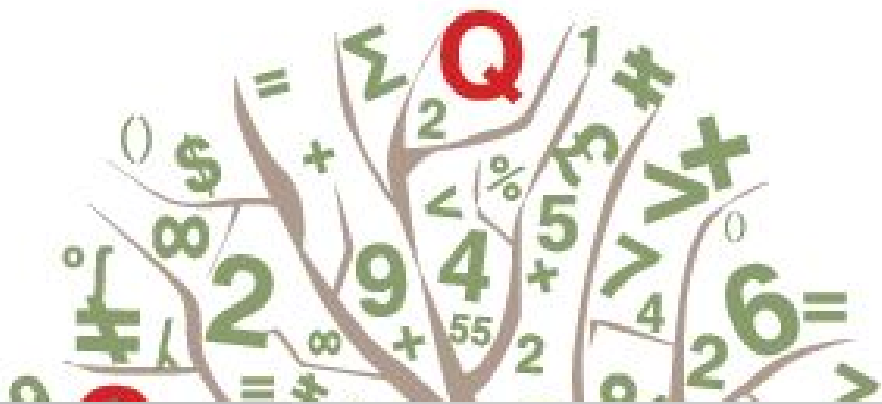
Q-Step: Week 5 Lecture

Bivariate Relationships

Spyros Kosmidis

Oxford

February 11, 2022



Roadmap

Previously

- Research Design
- Concepts and Measurement
- Descriptive Statistics and Visualization

Today

- Bivariate Relationships
  - Conditional means
  - Correlation
  - OLS regression

Next Week

- Multivariate OLS regression

Conditional Means

Q: What was the mean level of turnout in 2017 in Brexit and Remain constituencies?

Conditional Means

Q: What was the mean level of turnout in 2017 in Brexit and Remain constituencies?

Conditional Means

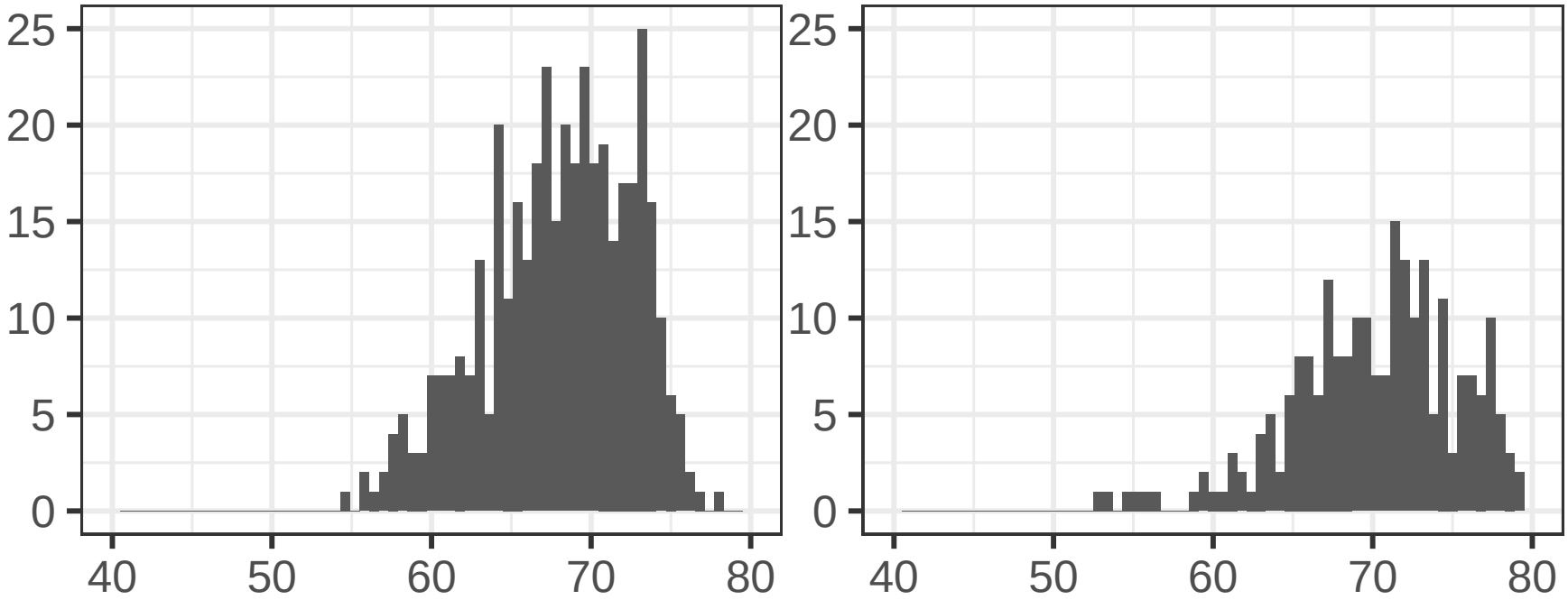
Q: What was the mean level of turnout in 2017 in Brexit and Remain constituencies?

Let's plot constituency turnout for Brexit and Remain constituencies:

Conditional Means

Q: What was the mean level of turnout in 2017 in Brexit and Remain constituencies?

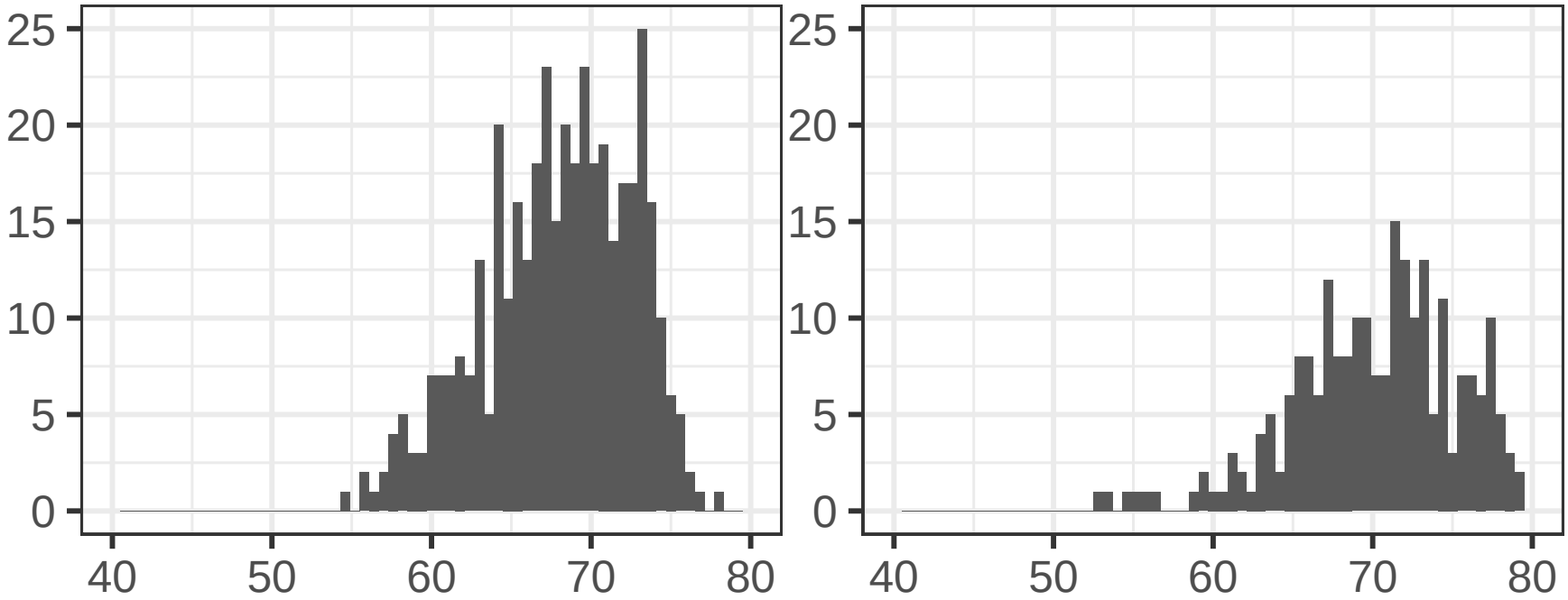
Let's plot constituency turnout for Brexit and Remain constituencies:



Conditional Means

Q: What was the mean level of turnout in 2017 in Brexit and Remain constituencies?

Let's plot constituency turnout for Brexit and Remain constituencies:

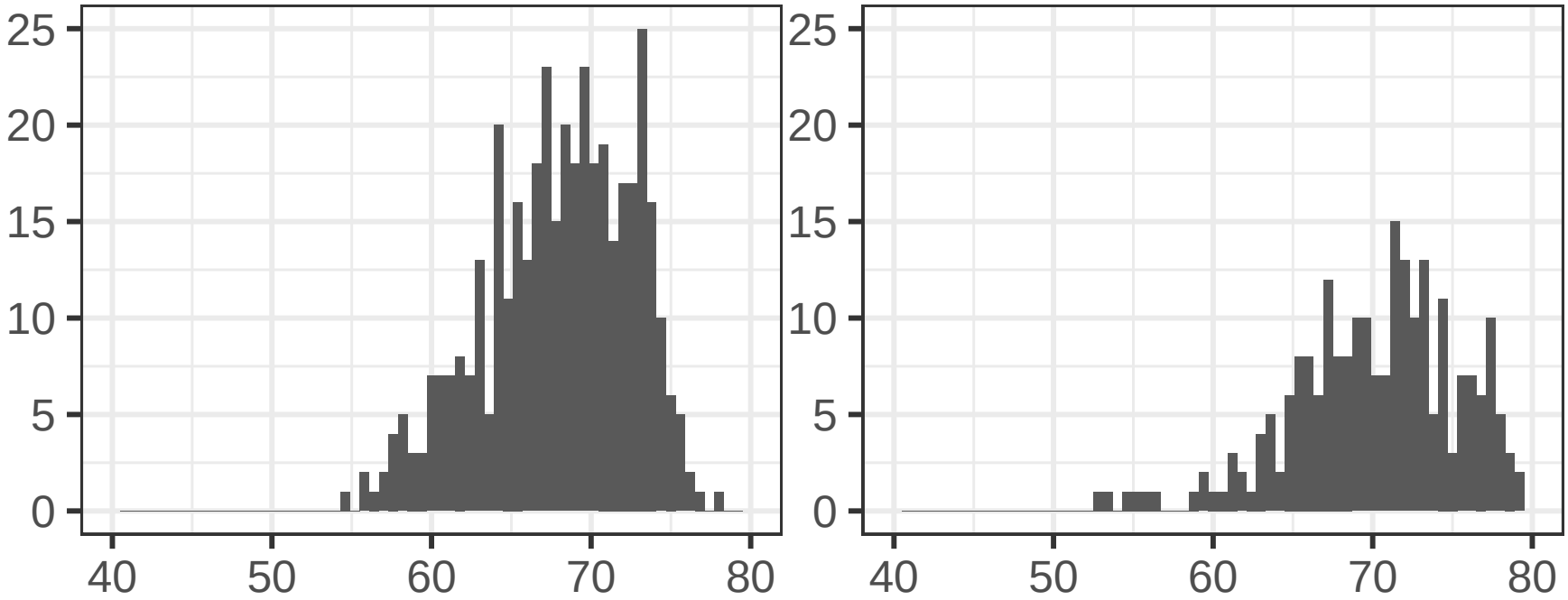


What do you observe (if anything)?

Conditional Means

Q: What was the mean level of turnout in 2017 in Brexit and Remain constituencies?

Let's plot constituency turnout for Brexit and Remain constituencies:



What do you observe (if anything)?

The mean level of turnout for all regions was 68.75%

Turnout in Brexit and Remain Areas (2017)

```
summary(brseat$Turnout[brseat$Year1==2017])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   53.02   65.42   69.16   68.75   72.39   79.52
```

```
mean(brseat$Turnout[brseat$Year1==2017 & brseat$remain==0])
```

```
## [1] 67.96286
```

```
mean(brseat$Turnout[brseat$Year1==2017 & brseat$remain==1])
```

```
## [1] 70.13509
```

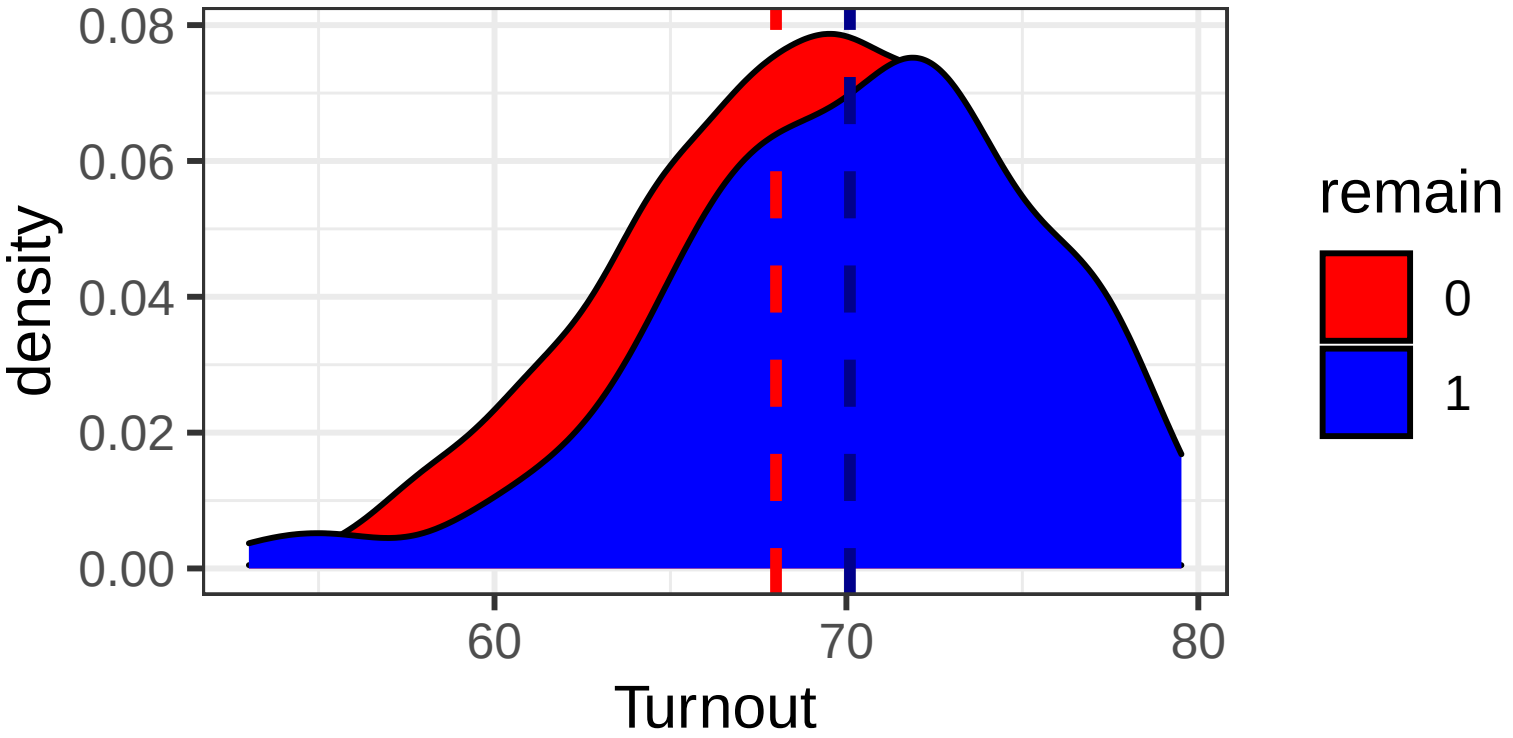
Turnout in Brexit and Remain Areas (2017)

```
median(brseat$Turnout[brseat$Year1==2017 & brseat$remain==0])
```

```
## [1] 68.50045
```

```
median(brseat$Turnout[brseat$Year1==2017 & brseat$remain==1])
```

```
## [1] 70.85917
```



Did Brexit Increase Turnout in 2017?

Did Brexit Increase Turnout in 2017?

```
mean(brseat$Turnout[brseat$Year1==2017])
```

```
## [1] 68.74995
```

```
mean(brseat$Turnout[brseat$Year1==2015])
```

```
## [1] 66.38139
```

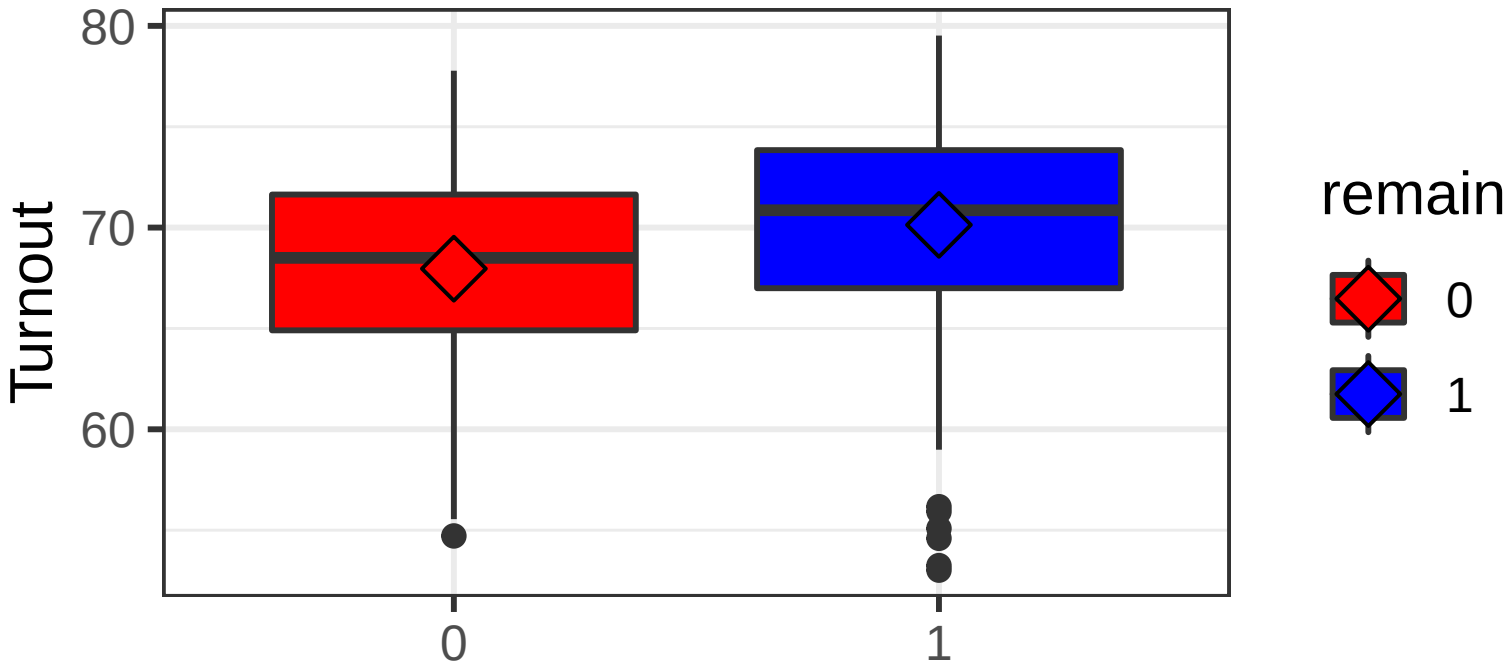
Did Brexit Increase Turnout in 2017?

```
mean(brseat$Turnout[brseat$Year1==2017])
```

```
## [1] 68.74995
```

```
mean(brseat$Turnout[brseat$Year1==2015])
```

```
## [1] 66.38139
```



Did Brexit Increase Turnout in 2017?

Did Brexit Increase Turnout in 2017?

```
mean(brseat$Turnout[brseat$Year1==2017 & brseat$remain==0])
```

```
## [1] 67.96286
```

```
mean(brseat$Turnout[brseat$Year1==2017 & brseat$remain==1])
```

```
## [1] 70.13509
```

Did Brexit Increase Turnout in 2017?

```
mean(brseat$Turnout[brseat$Year1==2017 & brseat$remain==0])
```

```
## [1] 67.96286
```

```
mean(brseat$Turnout[brseat$Year1==2017 & brseat$remain==1])
```

```
## [1] 70.13509
```

```
mean(brseat$Turnout[brseat$Year1==2015 & brseat$remain==0])
```

```
## [1] 65.28172
```

```
mean(brseat$Turnout[brseat$Year1==2015 & brseat$remain==1])
```

```
## [1] 68.3166
```

Conditional Means

- Simple and intuitive way to understand data
- Extremely easy to calculate and visualise
- Powerful but constrained by the categorical conditions
- What if you have continuous variables that might be related?

Correlation

Correlation

- Correlation

Correlation

- Correlation
  - It shows direction and strength

Correlation

- Correlation
  - It shows direction and strength
- But bad for predictions

Correlation

- Correlation
  - It shows direction and strength
- But bad for predictions
- Only Bivariate

Correlation

- Correlation
  - It shows direction and strength
- But bad for predictions
- Only Bivariate
- The coefficient runs from -1 to $+1$
  - 0 means no correlation
  - 1 means perfect positive correlation
  - -1 means perfect negative correlation

Correlation

- Correlation
  - It shows direction and strength
- But bad for predictions
- Only Bivariate
- The coefficient runs from -1 to +1
  - 0 means no correlation
  - 1 means perfect positive correlation
  - -1 means perfect negative correlation
- The software gives you two important measures
  - A confidence interval (i.e. a range of correlation values)
  - A p-value, i.e. a probability that the correlation is random

Correlation: Example

Correlation: Example

- Is there a correlation between educational qualifications and Brexit?

Correlation: Example

- Is there a correlation between educational qualifications and Brexit?

```
##
##  Pearson's product-moment correlation
##
## data:  brexit$leave and brexit$noqual
## t = 11.697, df = 377, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4380723 0.5862924
## sample estimates:
##       cor
## 0.5160348
```

Correlation: Example

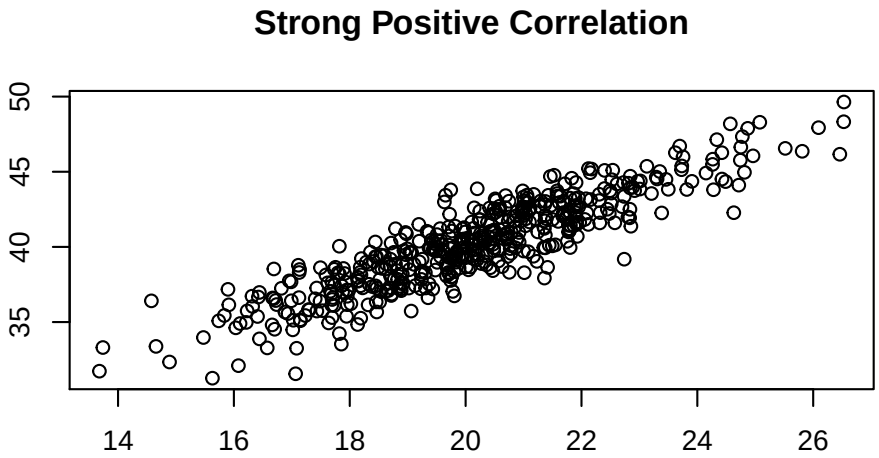
- Is there a correlation between educational qualifications and Brexit?

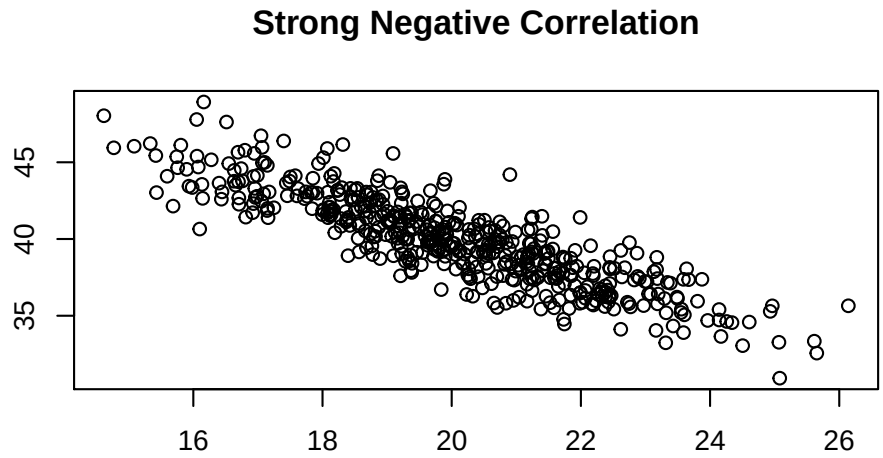
```
##
##  Pearson's product-moment correlation
##
## data:  brexit$leave and brexit$noqual
## t = 11.697, df = 377, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4380723 0.5862924
## sample estimates:
##       cor
## 0.5160348
```

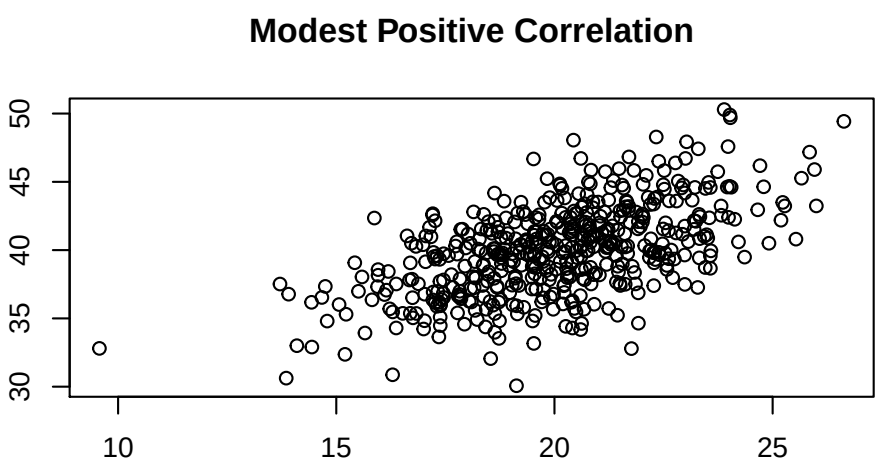
$Cor(x, y) = \frac{Cov(x,y)}{sd(x)sd(y)}$ where $Cov(x, y) = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n-1}$

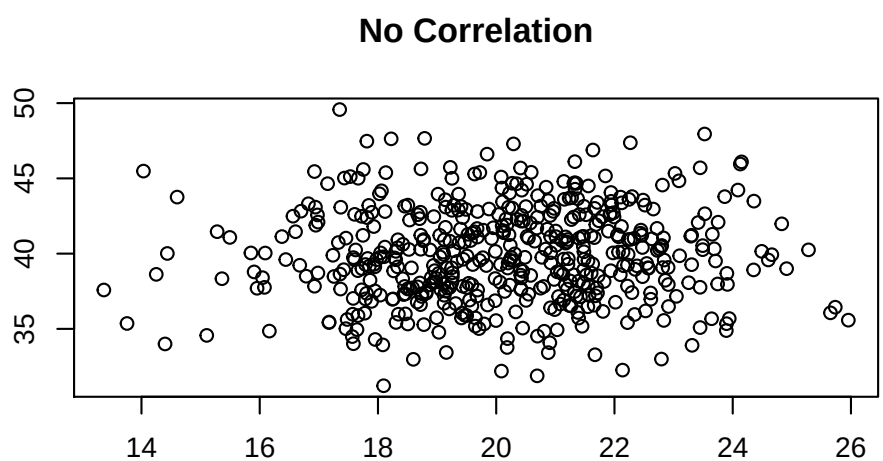
- This coefficient suggests that the higher the proportion of residents in a given locality without ANY educational qualifications the larger the vote share for Leave

Interpretation









OLS regression

OLS regression

- Regression
  - Good for Prediction and Explanation
  - Both Bivariate and Multivariate
  - The most widely used technique

OLS regression

- Regression
  - Good for Prediction and Explanation
  - Both Bivariate and Multivariate
  - The most widely used technique
- Key Logic

$$Y_i = \alpha + \beta X_i$$

OLS regression

- Regression
  - Good for Prediction and Explanation
  - Both Bivariate and Multivariate
  - The most widely used technique
- Key Logic

$$Y_i = \alpha + \beta X_i$$

We try to find the best line β that minimizes the amount of 'error' in the predictions.

OLS regression

- Regression
  - Good for Prediction and Explanation
  - Both Bivariate and Multivariate
  - The most widely used technique
- Key Logic

$$Y_i = \alpha + \beta X_i$$

We try to find the best line β that minimizes the amount of 'error' in the predictions.

We will cover the "error" part shortly.

OLS regression: Example

OLS regression: An example

The OLS output in R

```
##
## Call:
## lm(formula = brexit$leave ~ brexit$noqual)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.855  -3.593   1.971   5.958  24.182
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   29.33773    2.08661   14.06   <2e-16 ***
## brexit$noqual  1.04234    0.08911   11.70   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.945 on 377 degrees of freedom
## Multiple R-squared:  0.2663, Adjusted R-squared:  0.2643
## F-statistic: 136.8 on 1 and 377 DF,  p-value: < 2.2e-16
```

The OLS output in Papers

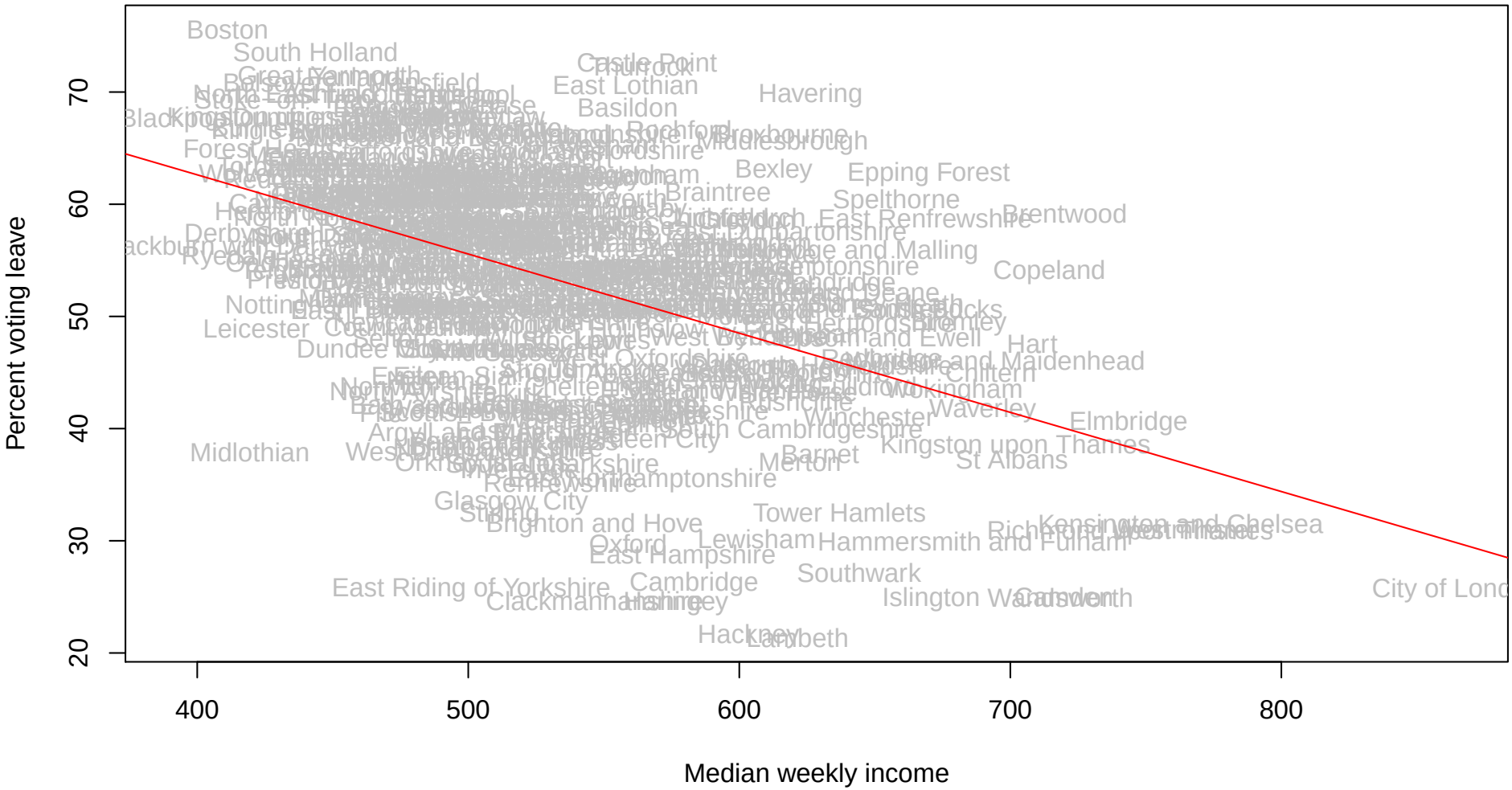
Table 1:

	Dependent variable:
	leave
noqual	1.042^{***}
	(0.089)
Constant	29.338^{***}
	(2.087)
Observations	379
R^2	0.266
Adjusted R^2	0.264
Residual Std. Error	8.945 (df = 377)
F Statistic	136.828^{***} (df = 1; 377)

OLS Regression: A Second Example

```
##
## Call:
## lm(formula = brexit$leave ~ brexit$income)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.804  -5.471   1.452   5.837  23.010
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   90.853597   3.644753   24.93   <2e-16 ***
## brexit$income -0.070581   0.006767  -10.43   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.289 on 360 degrees of freedom
##   (17 observations deleted due to missingness)
## Multiple R-squared:  0.232,  Adjusted R-squared:  0.2299
## F-statistic: 108.8 on 1 and 360 DF,  p-value: < 2.2e-16
```

OLS Regression: A Second Example

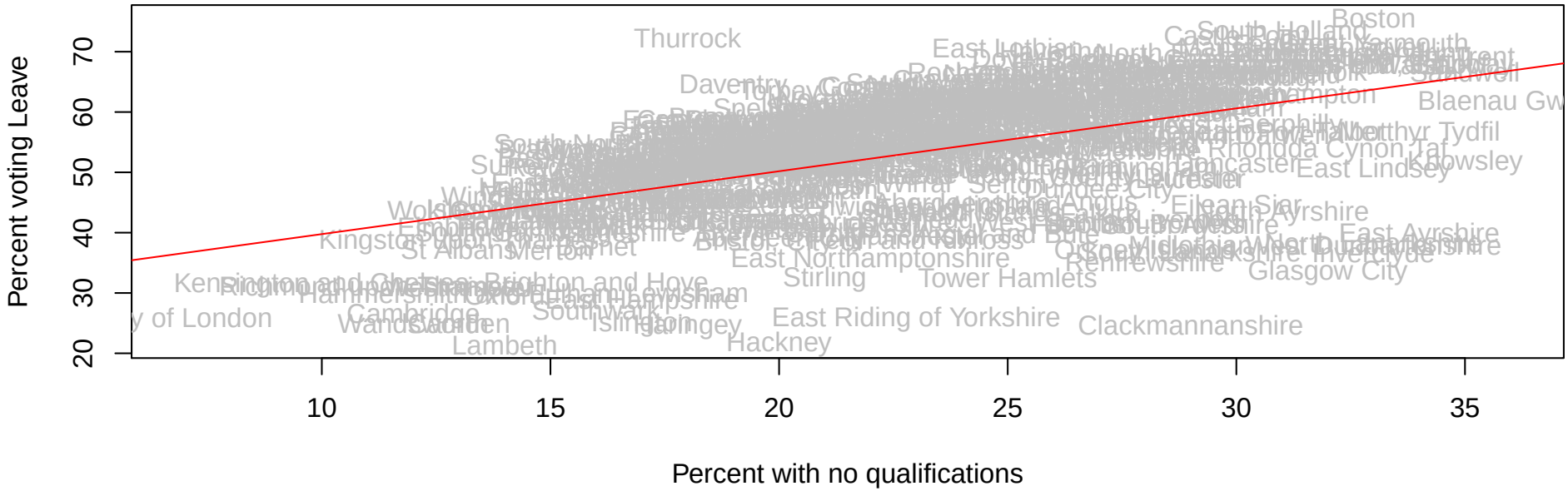


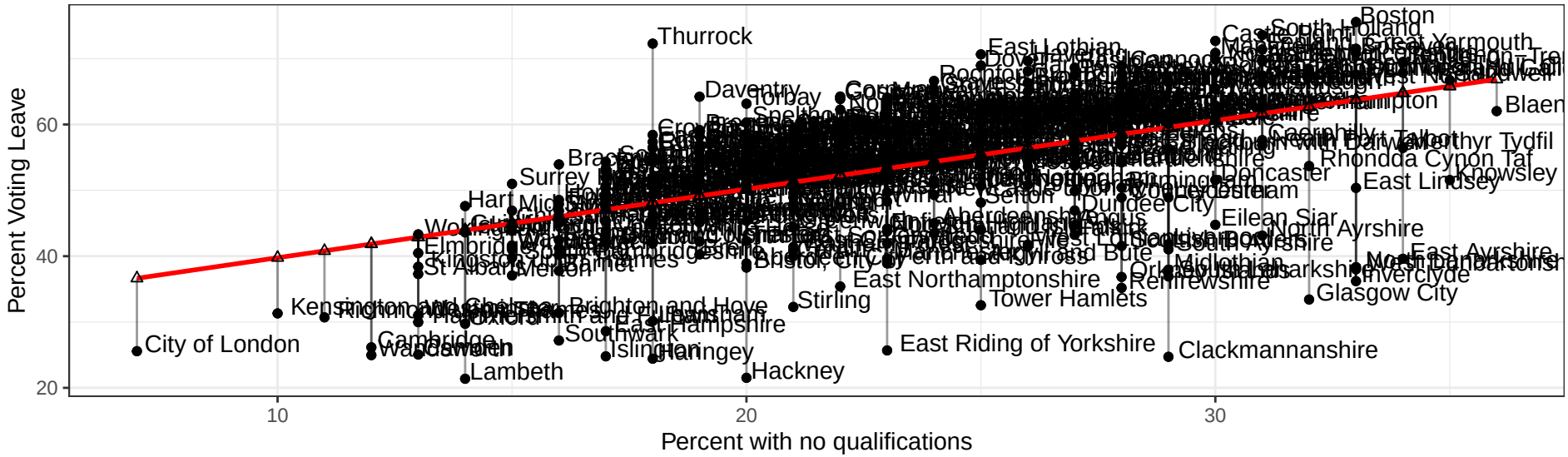
OLS Regression: A Second Example

Table 2:

	Dependent variable:
	leave
income	-0.071^{***}
	(0.007)
Constant	90.854^{***}
	(3.645)
Observations	362
R^2	0.232
Adjusted R^2	0.230
Residual Std. Error	9.289 (df = 360)
F Statistic	108.780^{***} (df = 1; 360)

Back to Example 1, How to fit the line





Basic Interpretation of OLS

Basic Interpretation of OLS

Recall the linear equation;

Basic Interpretation of OLS

Recall the linear equation;

$$Y_i = \alpha + \beta X_i$$

Basic Interpretation of OLS

Recall the linear equation;

$$Y_i = \alpha + \beta X_i$$

- How do we interpret the OLS coefficient?
  - A unit increase in X produce a coefficient increase in Y

Basic Interpretation of OLS

Recall the linear equation;

$$Y_i = \alpha + \beta X_i$$

- How do we interpret the OLS coefficient?
  - A unit increase in X produce a coefficient increase in Y
- How do we interpret the constant?
  - The constant gives the mean value of Y when X equals to 0

Basic Interpretation of OLS

Recall the linear equation;

$$Y_i = \alpha + \beta X_i$$

- How do we interpret the OLS coefficient?
  - A unit increase in X produce a coefficient increase in Y
- How do we interpret the constant?
  - The constant gives the mean value of Y when X equals to 0
- What is the standard error of a coefficient?
  - It is a measure of dispersion of our estimates (more in week 7)

Basic Interpretation of OLS

Recall the linear equation;

$$Y_i = \alpha + \beta X_i$$

- How do we interpret the OLS coefficient?
  - A unit increase in X produce a coefficient increase in Y
- How do we interpret the constant?
  - The constant gives the mean value of Y when X equals to 0
- What is the standard error of a coefficient?
  - It is a measure of dispersion of our estimates (more in week 7)
- What is a p-value?
  - The p-value is our test of statistical significance (In practice, if $p > 0.05$ then there is no statistically significant effect)
  - In the Week 7 lecture I will give you a more precise of interpretation and a clearer discussion of where it comes from.

Next week

- We will be moving to multivariate relationships
- I will begin with a recap of what we did today and then expand on multivariate OLS
- We will be interpreting more models and we will also look at additional regression checks (e.g model fit)
- Until then, make sure you work on the lab worksheets!

Thank You