Q-Step: Week 6 Lecture

Multivariate Relationships

Spyros Kosmidis

Oxford

December 21, 2022

Roadmap

Previously

- Research Design
- Concepts and Measurement
- Descriptive Statistics and Visualization
- Bivariate Relationships
  - Conditional means
  - Correlation
  - Bivariate regression

Today

- Multivariate OLS regression

Correlation Analysis!

Correlation Analysis!

- Correlation

Correlation Analysis!

- Correlation
  - It shows direction and strength

Correlation Analysis!

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

Correlation Analysis!

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

Correlation Analysis!

- 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 Analysis!

- 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
```

Bivariate OLS recap!

- Key Logic

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

Bivariate OLS recap!

- 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.

Bivariate OLS recap!

- 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.

$$\Sigma\epsilon^2 = 0$$

Bivariate OLS recap!

- 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.

$$\Sigma\epsilon^2 = 0$$

We are looking for the unknowns (α, β) that satisfy the above

Bivariate OLS recap!

- 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.

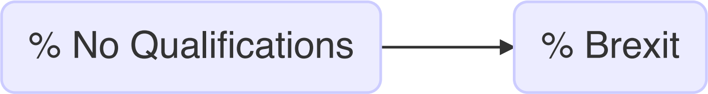
$$\Sigma\epsilon^2 = 0$$

We are looking for the unknowns (α, β) that satisfy the above

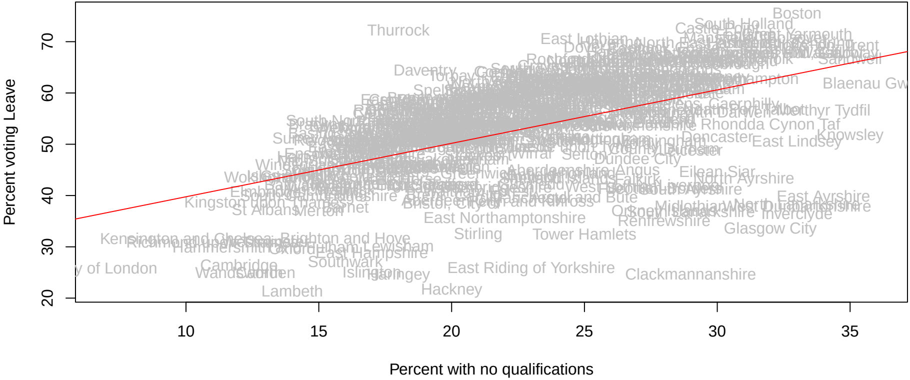
Let's go back to our examples from last week!

Our theory

Our theory



Bivariate OLS



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
```

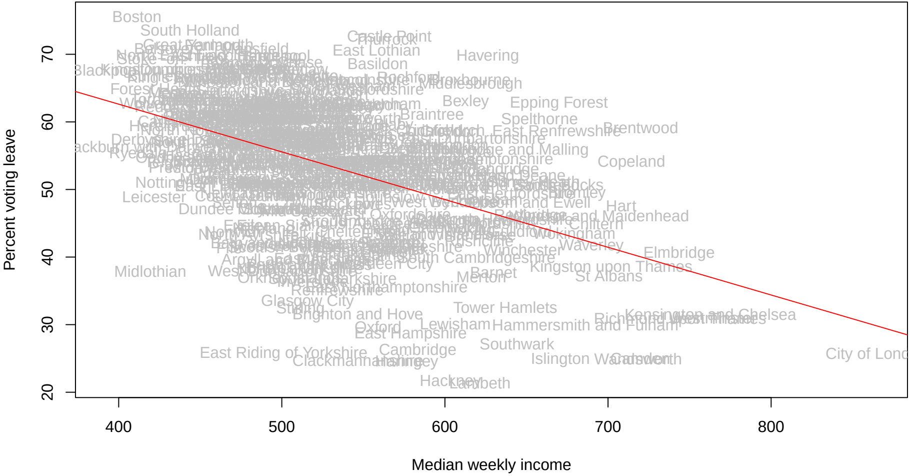
New theory

New theory

Bivariate 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
```

Bivariate OLS Regression: A Second Example



Basic Interpretation of OLS

- How do we interpret the OLS coefficient?
  - A unit increase in X predicts a coefficient increase in Y
  - In our case, if we increase median weekly income by a pound we get a 0.07 decrease in the percentage voting leave

Basic Interpretation of OLS

- How do we interpret the OLS coefficient?
  - A unit increase in X predicts a coefficient increase in Y
  - In our case, if we increase median weekly income by a pound we get a 0.07 decrease in the percentage voting leave
- How do we interpret the constant?
  - The constant gives the mean value of Y when X equals to 0
  - When a local area has zero median weekly income, then the average % voting leave is 90.8%
  - Remember that often times there is no 0 observation in our data sets!

Basic Interpretation of OLS

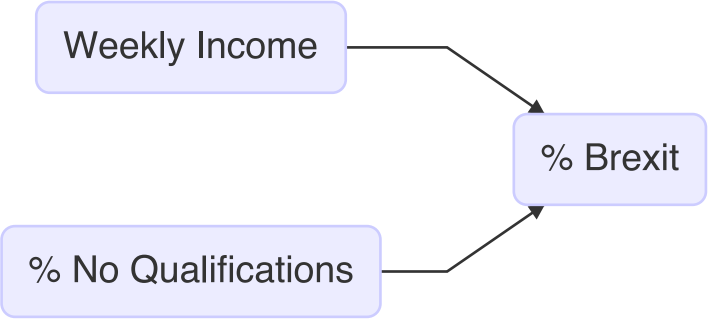
- How do we interpret the OLS coefficient?
  - A unit increase in X predicts a coefficient increase in Y
  - In our case, if we increase median weekly income by a pound we get a 0.07 decrease in the percentage voting leave
- How do we interpret the constant?
  - The constant gives the mean value of Y when X equals to 0
  - When a local area has zero median weekly income, then the average % voting leave is 90.8%
  - Remember that often times there is no 0 observation in our data sets!
- How do we assess statistical significance?
  - The p-value is our test of statistical significance (In practice, if $p > 0.05$ then there is no statistically significant effect)
  - In our regression all our effects are statistically significant!

Basic Interpretation of OLS

- How do we interpret the OLS coefficient?
  - A unit increase in X predicts a coefficient increase in Y
  - In our case, if we increase median weekly income by a pound we get a 0.07 decrease in the percentage voting leave
- How do we interpret the constant?
  - The constant gives the mean value of Y when X equals to 0
  - When a local area has zero median weekly income, then the average % voting leave is 90.8%
  - Remember that often times there is no 0 observation in our data sets!
- How do we assess statistical significance?
  - The p-value is our test of statistical significance (In practice, if $p > 0.05$ then there is no statistically significant effect)
  - In our regression all our effects are statistically significant!
- Future Steps
  - What if there is a second variable that might influence our outcome, but might also influence how our main X (e.g. weekly income) relates to Y?

A more complex theory

A more complex theory



F(% Brexit) –> F –> M –> A –>

Multivariate OLS

- With more than one predictor variables we fit multivariate regressions

Multivariate OLS

- With more than one predictor variables we fit multivariate regressions
- The equation for multivariate OLS is:

Multivariate OLS

- With more than one predictor variables we fit multivariate regressions
- The equation for multivariate OLS is:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots \beta_k X_k$$

- This time we estimate the partial correlation holding the other confounders constant
- The interpretation for e.g. β_1 is the same. Holding all other Xs constant, an increase in X_1 predicts a β_1 change in Y!

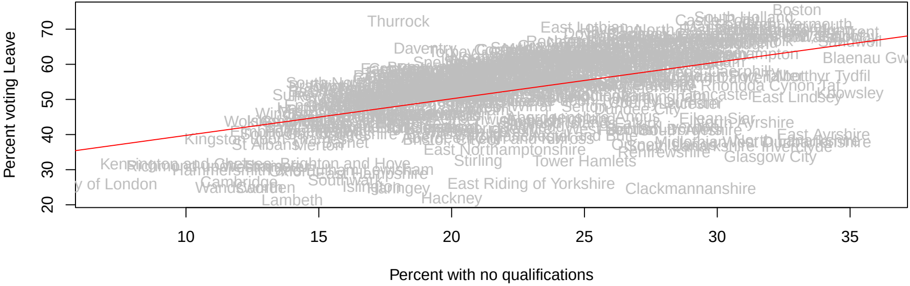
Multivariate OLS

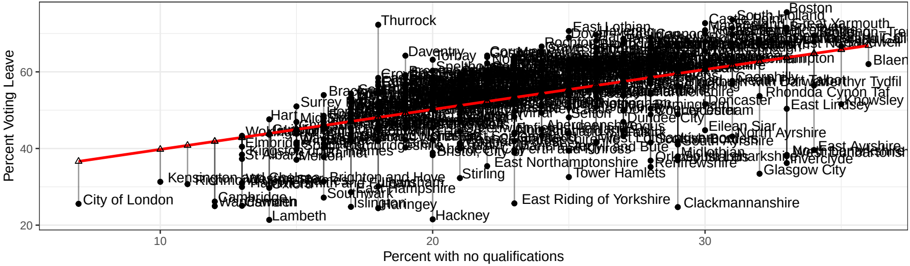
- With more than one predictor variables we fit multivariate regressions
- The equation for multivariate OLS is:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots \beta_k X_k$$

- This time we estimate the partial correlation holding the other confounders constant
- The interpretation for e.g. β_1 is the same. Holding all other Xs constant, an increase in X_1 predicts a β_1 change in Y!
- We still want to minimize the sum of the squared residuals

Bivariate Residuals





Estimating Multivariate Models

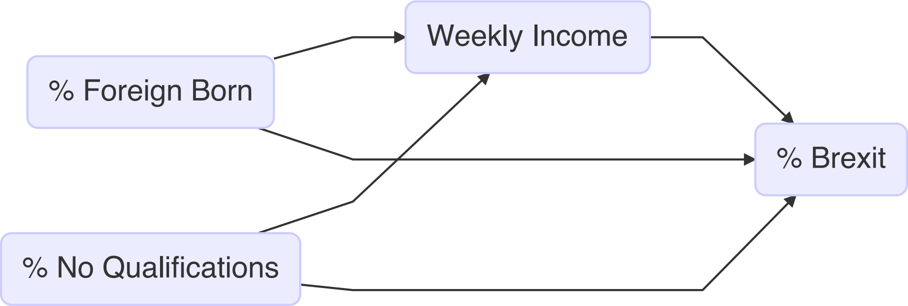
```
##
## Call:
## lm(formula = brexit$leave ~ brexit$income + brexit$noqual)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.518  -3.815   1.942   5.839  23.605
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   56.052694   6.866926   8.163 5.61e-15 ***
## brexit$income -0.036124   0.008729  -4.138 4.37e-05 ***
## brexit$noqual  0.721536   0.122667   5.882 9.28e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.883 on 359 degrees of freedom
##   (17 observations deleted due to missingness)
## Multiple R-squared:  0.2996, Adjusted R-squared:  0.2957
## F-statistic: 76.77 on 2 and 359 DF,  p-value: < 2.2e-16
```

Tidying up the output

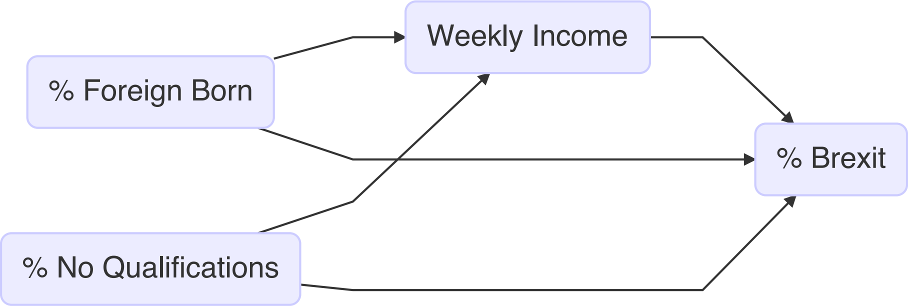
```
##
## ========================================
##                    Dependent variable:
##                ---------------------------
##                          leave
## ----------------------------------------
## income                 -0.036***
##                         (0.009)
##
## noqual                  0.722***
##                         (0.123)
##
## Constant               56.053***
##                         (6.867)
##
## ----------------------------------------
## Observations              362
## R2                       0.300
## Adjusted R2              0.296
## Residual Std. Error  8.883 (df = 359)
```

Social processes are complex!

Social processes are complex!



Social processes are complex!



- Don't read too much into the diagram, it is just to show you that social links can be complex!
- Still, multivariate models could tell us a lot about such a relationship!

How good is our model?

- We are some times interested in how well our model is performing
- The R^2 is a common fit statistic used by many
- It gives the proportion variance explained by the chosen model specification
- When one uses competing model specification, the R^2 and the adjusted-R^2 can be used
- In the past, people would place too much emphasis on model fit. I would encourage you to consider it, but don't go crazy about model fit.

How do we calculate the R^2?

How do we calculate the R^2?

The R^2 is defined as the ratio between the predicted and the actual variance?

How do we calculate the R^2?

The R^2 is defined as the ratio between the predicted and the actual variance?

$$R^2 = 1 - \frac{\Sigma(\hat{Y} - \bar{Y})^2}{\Sigma(Y - \bar{Y})^2}$$

How do we calculate the R^2?

The R^2 is defined as the ratio between the predicted and the actual variance?

$$R^2 = 1 - \frac{\Sigma(\hat{Y} - \bar{Y})^2}{\Sigma(Y - \bar{Y})^2}$$

- a key problem with the above equation and R^2 more generally is that adding more variables inflates the model fit.

How do we calculate the R^2?

The R^2 is defined as the ratio between the predicted and the actual variance?

$$R^2 = 1 - \frac{\Sigma(\hat{Y} - \bar{Y})^2}{\Sigma(Y - \bar{Y})^2}$$

- a key problem with the above equation and R^2 more generally is that adding more variables inflates the model fit.
- A penalty for additional parameters can be of help

How do we calculate the R^2?

The R^2 is defined as the ratio between the predicted and the actual variance?

$$R^2 = 1 - \frac{\Sigma(\hat{Y} - \bar{Y})^2}{\Sigma(Y - \bar{Y})^2}$$

- a key problem with the above equation and R^2 more generally is that adding more variables inflates the model fit.
- A penalty for additional parameters can be of help

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

where n is the number of observations and p is the number of parameters included in the model specification

Putting all models in one table

```
##
## ===================================================================================================
##                                              Dependent variable:
##                         ---------------------------------------------------------------------------
##                                                     leave
##                                    (1)                      (2)                      (3)
## ---------------------------------------------------------------------------------------------------
## income                        -0.071***                                         -0.036***
##                                (0.007)                                           (0.009)
##
## noqual                                                  1.042***                0.722***
##                                                         (0.089)                  (0.123)
##
## Constant                      90.854***                29.338***                56.053***
##                                (3.645)                  (2.087)                  (6.867)
##
## ---------------------------------------------------------------------------------------------------
## Observations                     362                      379                      362
## R2                              0.232                    0.266                    0.300
## Adjusted R2                     0.230                    0.264                    0.296
## Residual Std. Error      9.289 (df = 360)         8.945 (df = 377)         8.883 (df = 359)
## F Statistic          108.780*** (df = 1; 360) 136.828*** (df = 1; 377) 76.765*** (df = 2; 359)
## ===================================================================================================
## Note:                                                          *p<0.1; **p<0.05; ***p<0.01
```

Summary

Today

- We learned about Multivariate OLS
- It is the foundation of the vast majority of analyses in the social sciences
- It allows to test multiple hypotheses
- And make conditional predictions about continuous dependent variables
- We also talked about model fit
- What is left to wrap up OLS modeling?
- Uncertainty and Significance

Next Week

- A good overview of statistical inference
- Next week's lecture will help you grow your confidence in estimating regressions
- Some aspects of inference require a *leap of faith*, but most of it is straight forward!

Thank You

Click on the link below to donwload the data:

https://tinyurl.com/yc776n68