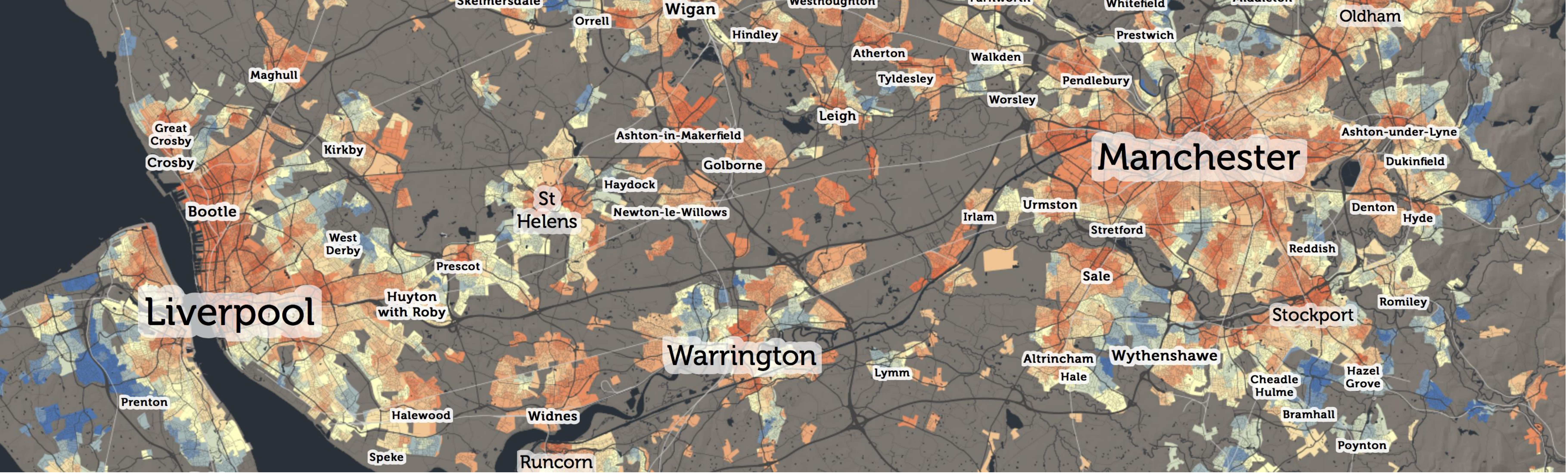


Multiple Linear Regression

ENVS225 Exploring the Social World

Gabriele Filomena
gfilo@liverpool.ac.uk

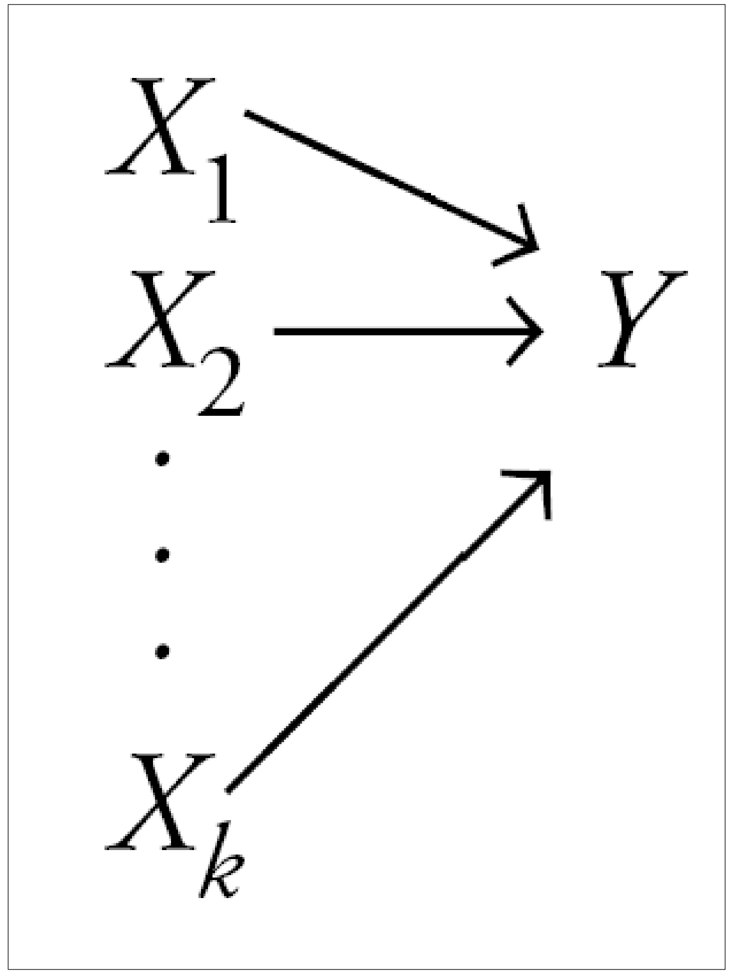
Learning goals

- Learn when to use multiple regression
- Learn how multiple regression extends simple linear regression
- Learn how to use multiple regression in real applications

Simple regression considers the relation between a single explanatory variable and response variable

$$X \to Y$$

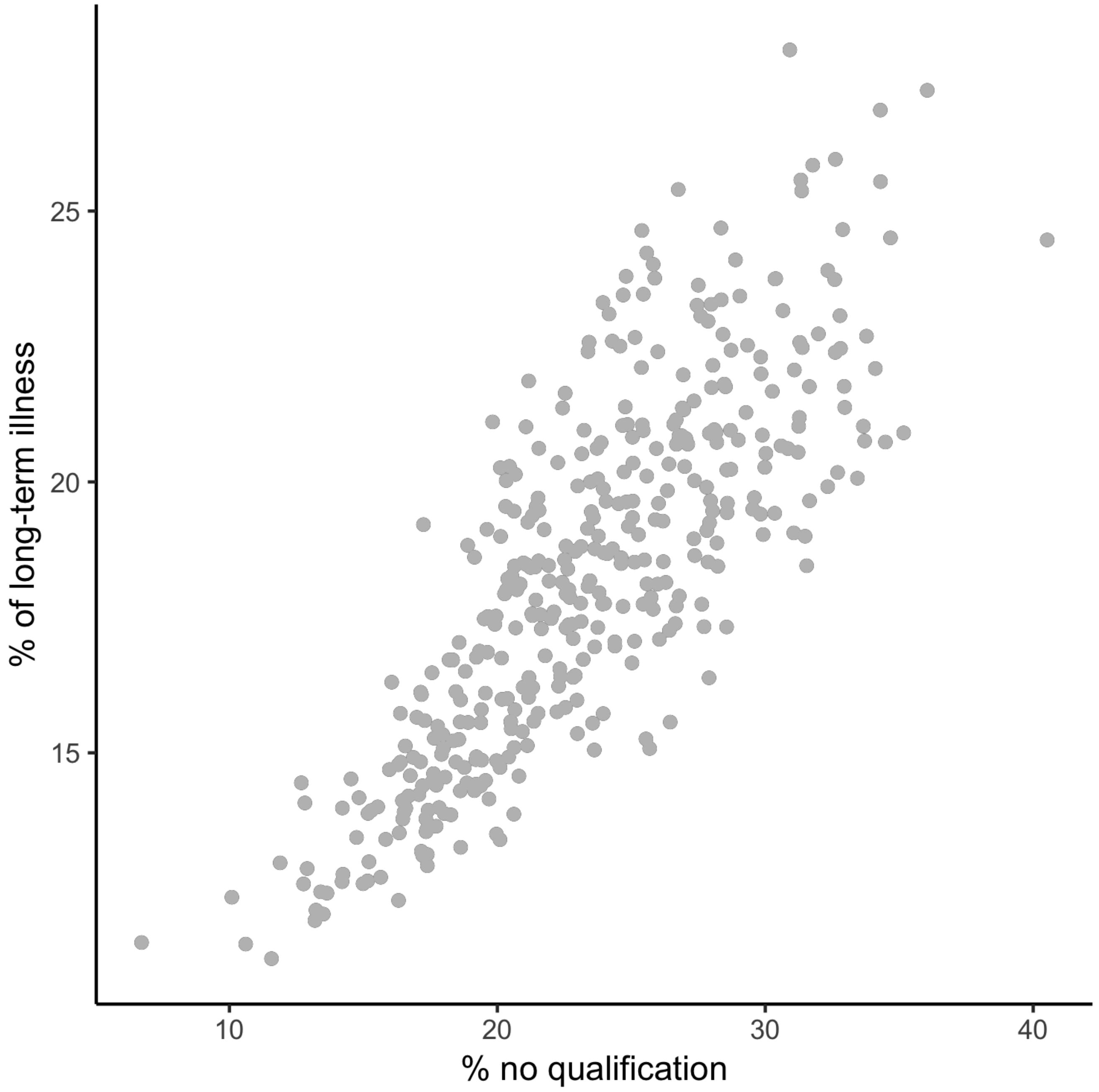
Multiple regression simultaneously considers the influence of multiple explanatory variables on a variable Y

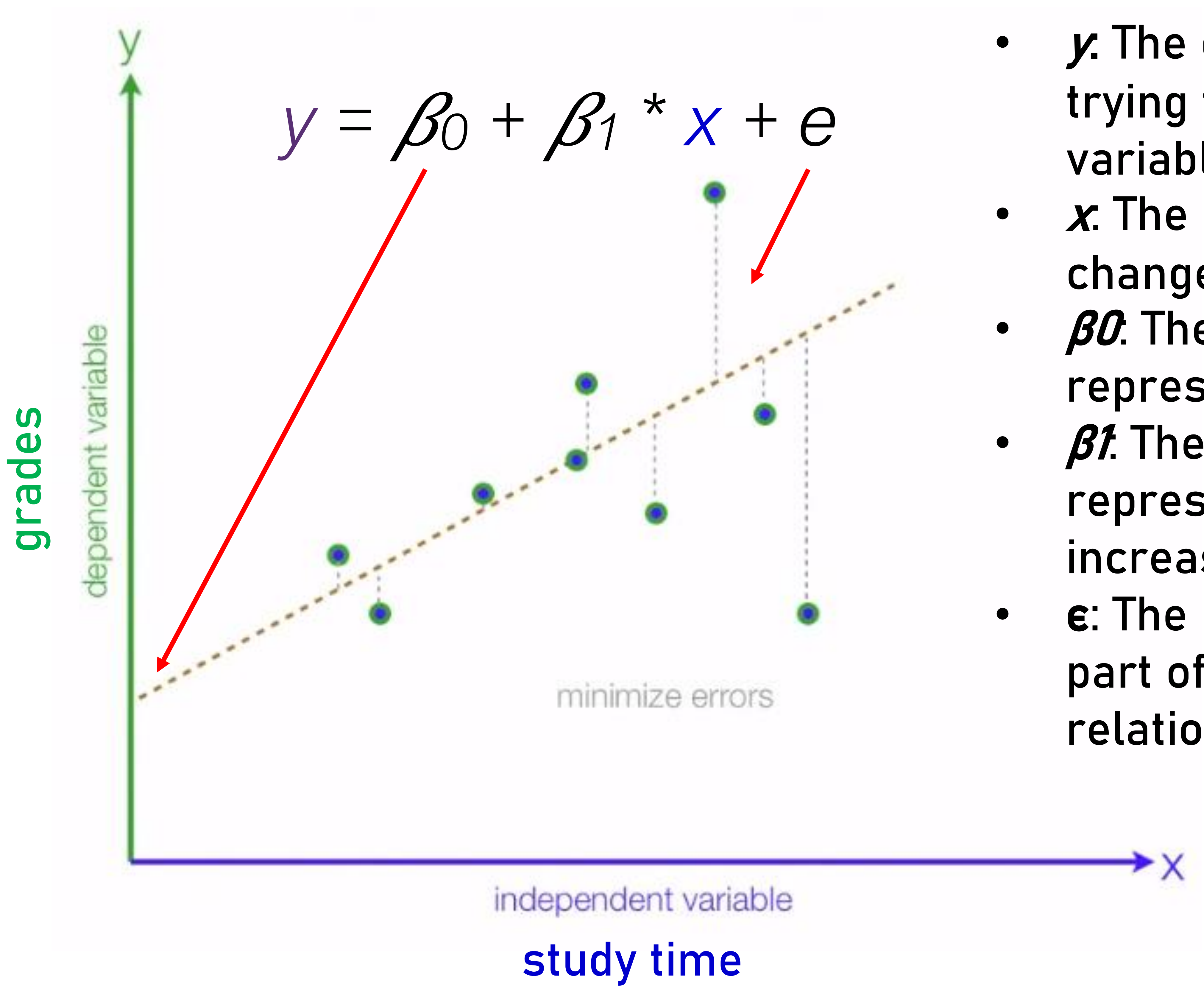


Simple Linear Regression Analysis

It is used to estimate the relationship between 2 continuous variables.

- It measures the relationship between two variables.
- It can tell the value of the Y at a certain value of X





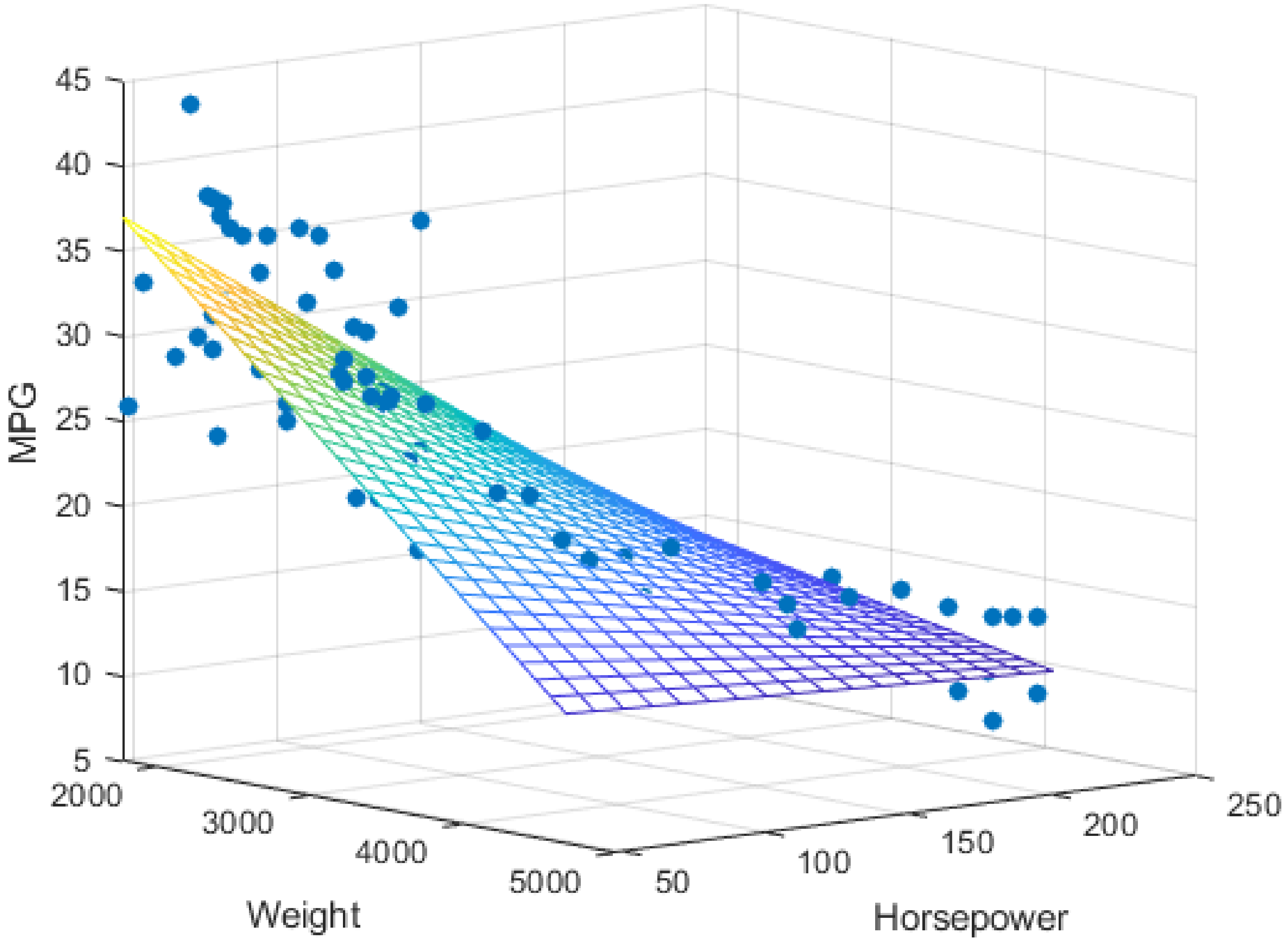
- ***y***: The dependent variable. The outcome we are trying to explain based on the independent variable.
- ***x***: The independent variable used to explain changes in the dependent variable.
- ***β_0***: The intercept of the regression line. It represents the expected value of *y* when *x* is 0.
- ***β_1***: The slope of the regression line. It represents the change in *y* for a one-unit increase in *x*.
- **ϵ**: The error term (or residual). It represents the part of *y* that cannot be explained by the linear relationship with *x*.

Multiple Regression Analysis

Method for studying the relationship between 1 dependent variable and 2+ independent variables.

Purposes:

- Explanation (our focus)
- Theory building
- Prediction



Simple	*vs*	Multiple
• One dependent variable Y predicted from one independent variable X.		• One dependent variable Y predicted from a set of independent variables ($X1, X2 \ldots Xk$).
• One regression coefficient.		• One regression coefficient for each independent variable.
• $\mathbf{r^2}$: proportion of variation in dependent variable Y predictable from X.		• $\boldsymbol{R^2}$: proportion of variation in Y predictable by set of independent variables (Xs).

Data Requirements

- One dependent variable
- Two or more independent variables (explanatory variables).
- Sample size: >= 50 (+ 10/20 per independent variables)
- Variables must be numerical or categorical converted into dummy variables (e.g. religion -> Catholic (y/n -> 1/0), Muslim (y/n -> 1/0 etc.) → Next week

Assumptions

- Independence: the scores of any particular subject are independent of the scores of all other subjects
- Normality: variables are normally distributed
- Homoscedasticity: the variances of the dependent variable for each of the possible combinations of the levels of the X variables are equal.
- Linearity: the relation between the dependent variable and the independent variable is linear when all the other independent variables are held constant.

The Model

Intercept

Slope coefficient 1

Slope coefficient 2

Slope coefficient 3

Dependent variable

Error

$$Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + \beta_3 X3 + \epsilon$$

Independent variable 1

Independent variable 2

Independent variable 3

Interpretation

Intercept (β_0)

- The estimated average value of Y when the value of the Xs is zero.

$$Y = \beta_0 + \cancel{\beta_1 X1}^{\,0} + \cancel{\beta_2 X2}^{\,0} + \cancel{\beta_3 X3}^{\,0} + \epsilon$$

Slope

- The estimated average change in Y for a one-unit change in X, when all other explanatory variables are held constant

$$Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + \beta_3 X3 + \epsilon$$

Example

$$Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + \beta_3 X3 + \epsilon$$

% of very bad health

% of males

% no qualification

% higher professionals

We are trying to find the *β*s

Output

```
Call:
lm(formula = pct_Very_bad_health ~ pct_No_qualifications + pct_Males +
    pct_Higher_manager_prof, data = census)

Residuals:
    Min      1Q  Median      3Q     Max
-0.4903 -0.1369 -0.0352  0.0983  0.7658

Coefficients:
                        Estimate Std. Error t value             Pr(>|t|)
(Intercept)              4.01799    0.88004    4.57            0.0000071 ***
pct_No_qualifications    0.05296    0.00591    8.96 < 0.0000000000000002 ***
pct_Males               -0.07392    0.01785   -4.14            0.0000440 ***
pct_Higher_manager_prof -0.01309    0.00494   -2.65               0.0084 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.213 on 327 degrees of freedom
Multiple R-squared:  0.61,  Adjusted R-squared:  0.607
F-statistic:  171 on 3 and 327 DF,  p-value: <0.0000000000000002
```

$$Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + \beta_3 X3 + \epsilon$$

% very bad health =

4.01 – 0.73*(% males) + 0.52*(%no qualification) – 0.013*(% higher professional)

Interpretation

- β_1 -> A 1%-point increase in the percentage of males decreases the percentage of population in very bad health by 0.73%
- β_2 -> A 1%-point increase in the percentage of no qualification population increases the percentage of population in very bad health by 0.52%
- β_0 -> If there were no local percentage of male, no qualification and higher professional population, the percentage of population in very bad health would be 4.01% (very unlikely scenario).

Units of measurement of the dependent and independent variables are important for interpretation purposes!

Hypothesis testing and Significance

For each variable $\boldsymbol{X_i}$, the null hypothesis:

- H_0 : There is no effect of $\boldsymbol{X_i}$ on $\boldsymbol{Y}$.

vs the alternative hypothesis:

- H_1 : There is an effect of $\boldsymbol{X_i}$ on $\boldsymbol{Y}$.

If the null hypothesis is rejected, there is an evidence that there is a significant relationship between $\boldsymbol{X_i}$ and Y.

T-test

- The t-test is performed as a hypothesis test to assess the significance of individual coefficients (or features) in the linear regression model.
- We look at the p-value of each coefficient β_i.
- If the p-value is less than 0.05, we reject the null hypothesis, otherwise, we do not.

What results should be reported?

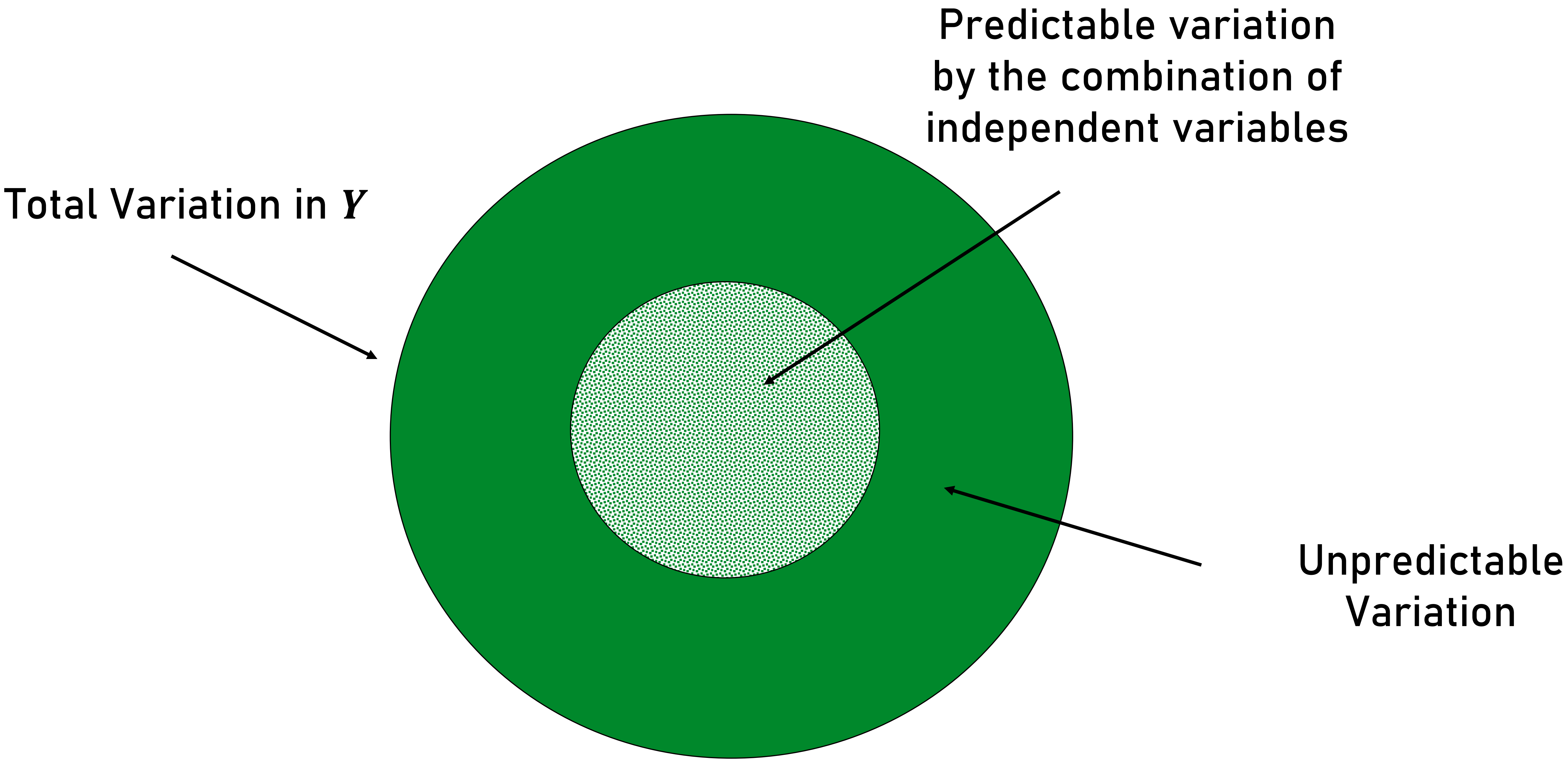
```
Coefficients:
                      Estimate Std. Error t value           Pr(>|t|)    
(Intercept)            4.01799    0.88004    4.57          0.0000071 ***
pct_No_qualifications  0.05296    0.00591    8.96 < 0.0000000000000002 ***
pct_Males             -0.07392    0.01785   -4.14          0.0000440 ***
pct_Higher_manager_prof -0.01309  0.00494   -2.65             0.0084 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.213 on 327 degrees of freedom
Multiple R-squared:  0.61,	Adjusted R-squared:  0.607 
F-statistic:  171 on 3 and 327 DF,  p-value: <0.0000000000000002
```

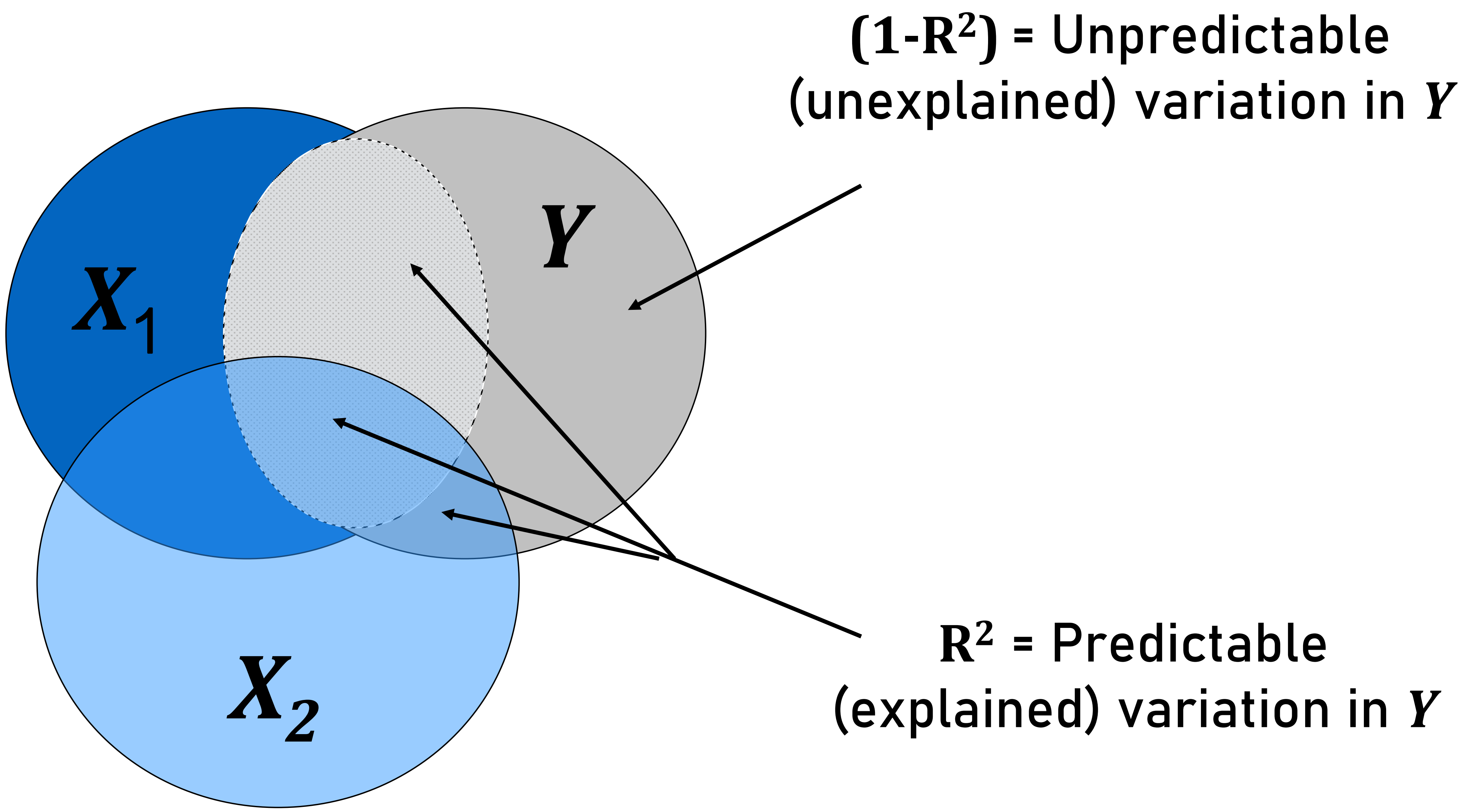
Coefficient of determination R^2

- Squared multiple correlation coefficient (R)
- Report (adjusted) R^2 instead of R
- Indicates the % of variance in the dependent variable explained by the combined effects of the IVs.
- The adjusted R^2 takes into account both the number of variables in the model and the sample size.

Explaining Variation: How much?



Proportion of Predictable and Unpredictable Variation



R^2 can be inflated by adding lots of predictors into the model even if most of these predictors are frivolous

Interpretation of R^2

- .00 = no linear relationship
- .10 = small
- .25 = moderate
- .50 = strong
- 1.00 = perfect linear relationship

Building Regression Models

- Simultaneous: all independent variables entered together
- Stepwise: independent variables entered according to some order (e.g. by size or correlation with dependent variable).

Week 8 Tasks

- Select a dataset: choose from those supplied in class, or one you have sourced yourself.
- Identify variables of interest.