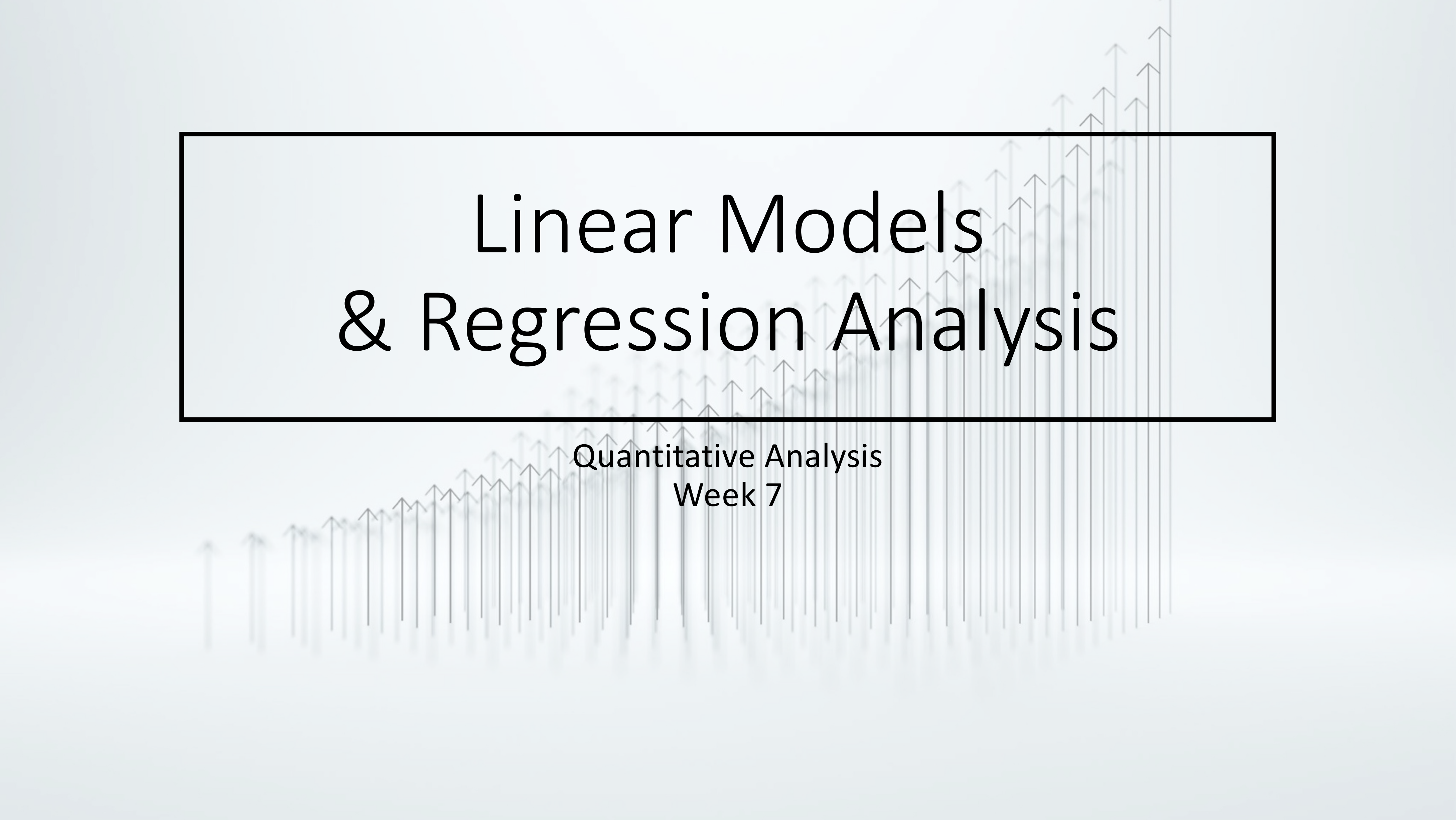


Linear Models
& Regression Analysis

Quantitative Analysis
Week 7

Week 5 Assignment

About Sample Sizes

- Several people suggested that the reason the T-tests didn’t match the expected mean (the income data from the National Statistics Office) was because the sample size was too small.
- 100 samples (50 of High School grads, 50 of Uni grads) is small – but it would be **extremely** unlikely to accidentally reject the expected mean with a truly random sample of this size.
- The sample size isn’t the problem: the problem is that the sample **was not random**.

When is a random sample not random?

- In the homework assignment question, it clearly stated that the samples of High School graduates and University graduates were randomly chosen.
- However, even if the samples are perfectly randomly chosen from those groups, they are not randomly chosen from the population overall!
- This sampling method could miss some groups entirely (what about people who didn't graduate High School?). Also, the sample size in each group is the same (n = 50), but are they truly equally sized groups in the overall population?
- These samples **were** random and correctly chosen at the original **level of observation** – but when you change that level to the whole population, the samples are now **heavily biased** and no longer reliable.
- This is an important principle: even if your data was perfectly correctly gathered, it may be biased and inappropriate **if you change the question you are asking**.

Review (1)

- Statistical tests are used to test simple hypotheses about our data.
  - e.g., “are these two samples from the same distribution?”
- They tell us if there is a **statistically significant** difference between two samples of data, or between a sample and an expected mean.
- The test we choose is based on the types of variables we are analysing: quantitative, categorical, etc.
- There are parametric and nonparametric tests.
  - Parametric tests make more assumptions about the data (e.g. the **assumption of normality**)
  - Nonparametric tests make fewer assumptions but have wider confidence intervals.

Review (2)

- **T-tests** are tests for quantitative outcome variables with two-category (binary) predictors – i.e. two-sample data.
  - You can also do a T-test on one sample and an expected mean value.
- **ANOVA** allows us to test quantitative outcome variables across more than two categories (multi-sample data).
  - It can only tell us if one or more of the samples is statistically different. To find out which ones are different, we need additional testing, such as Tukey's Honestly Significant Difference (**TukeyHSD**).
- **Chi-Square** tests are for categorical variables and categorical outcomes.
  - They tell us how likely it is that a sample was drawn from evenly distributed categories (goodness of fit), and whether two categorical variables are independent of each other (test of independence).
  - For samples with any small expected values (< 5), **Fisher's Exact Test** should be used.

Review (3)

- When you have two quantitative variables, you can find out if they are related to each other by calculating their **correlation coefficient**.
- This has a value from -1 (a perfect inverse correlation) to +1 (a perfect direct correlation), with 0 indicating absolutely no correlation.

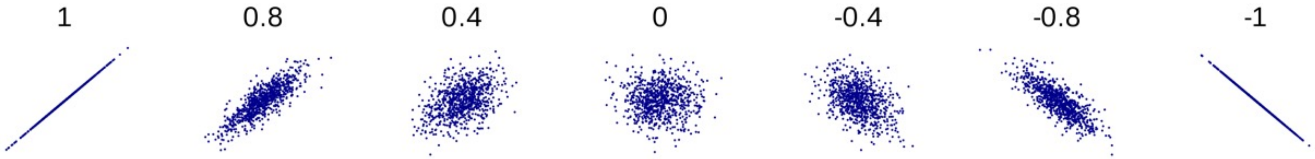
- Correlation is usually calculated using **Pearson's *r***,
    - This is a parametric test which assumes that the variables are **continuous** (i.e. they can take on any value, like age or salary, rather than being ranked, like responses on a five-point survey scale), and that the relationship is **linear** – so the best fit line on a scatter plot would be a straight line, not a curve.

- The nonparametric alternative is **Spearman's *rho***, which provides a better assessment of things like ranked data or variables with non-linear relationships.

Limitations of Correlation

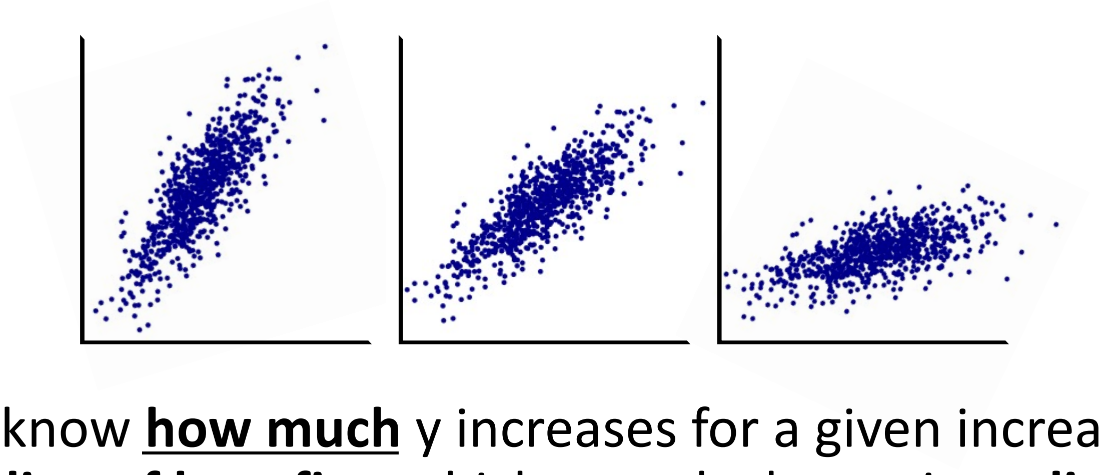
Correlation is a helpful tool for figuring out if two variables have a statistical relationship – but it doesn't fit every case and doesn't tell us everything we need to know.

It tells us the strength of the relationship – how much of *y* is predicted by *x* – but it doesn't tell us the **slope** of the line and misses some **complex relationships** entirely.



Line of Best Fit

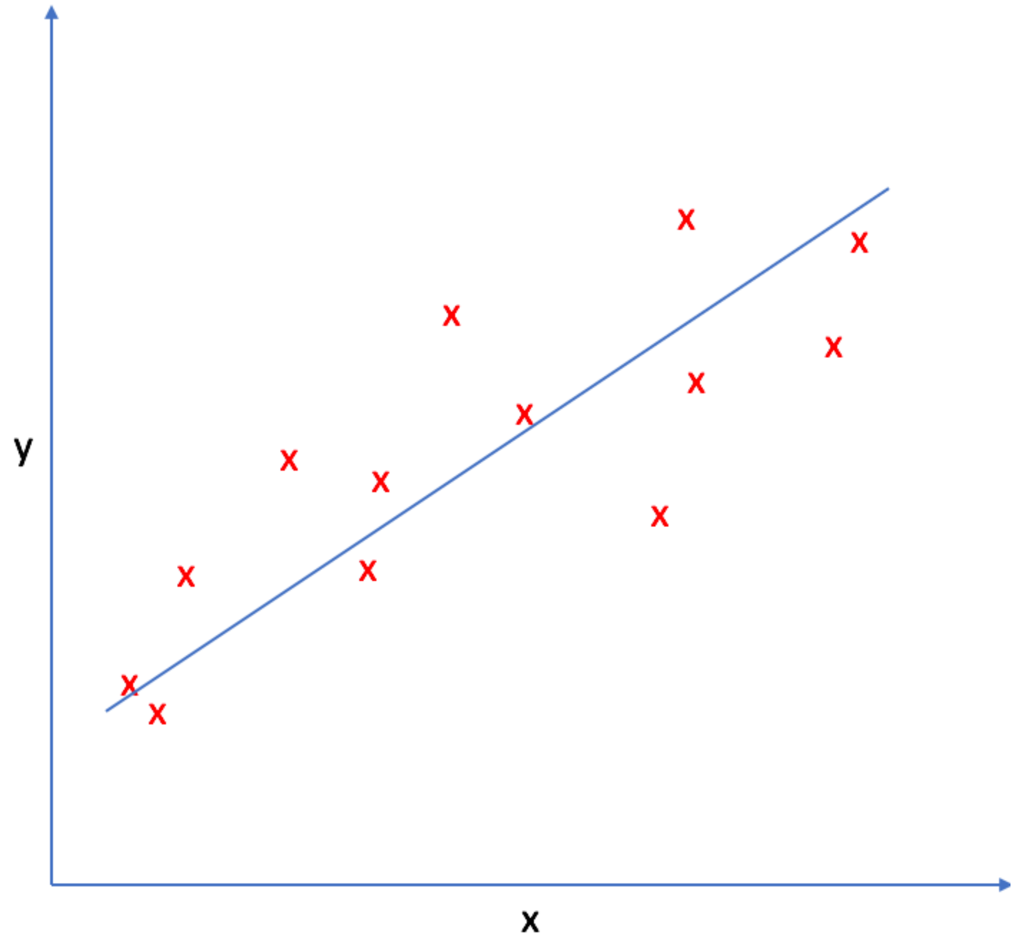
- Consider a bivariate (two variables) analysis where we find a correlation coefficient of 0.8.
- That's quite a strong correlation, so we know that <u>the value of x predicts the value of y quite well</u> – but it could refer to any of the graphs seen here.



- To actually know <u>**how much**</u> y increases for a given increase in x, we instead turn to the <u>**line of best fit**</u> – which we calculate using a <u>**linear regression**</u>.

Linear Regression

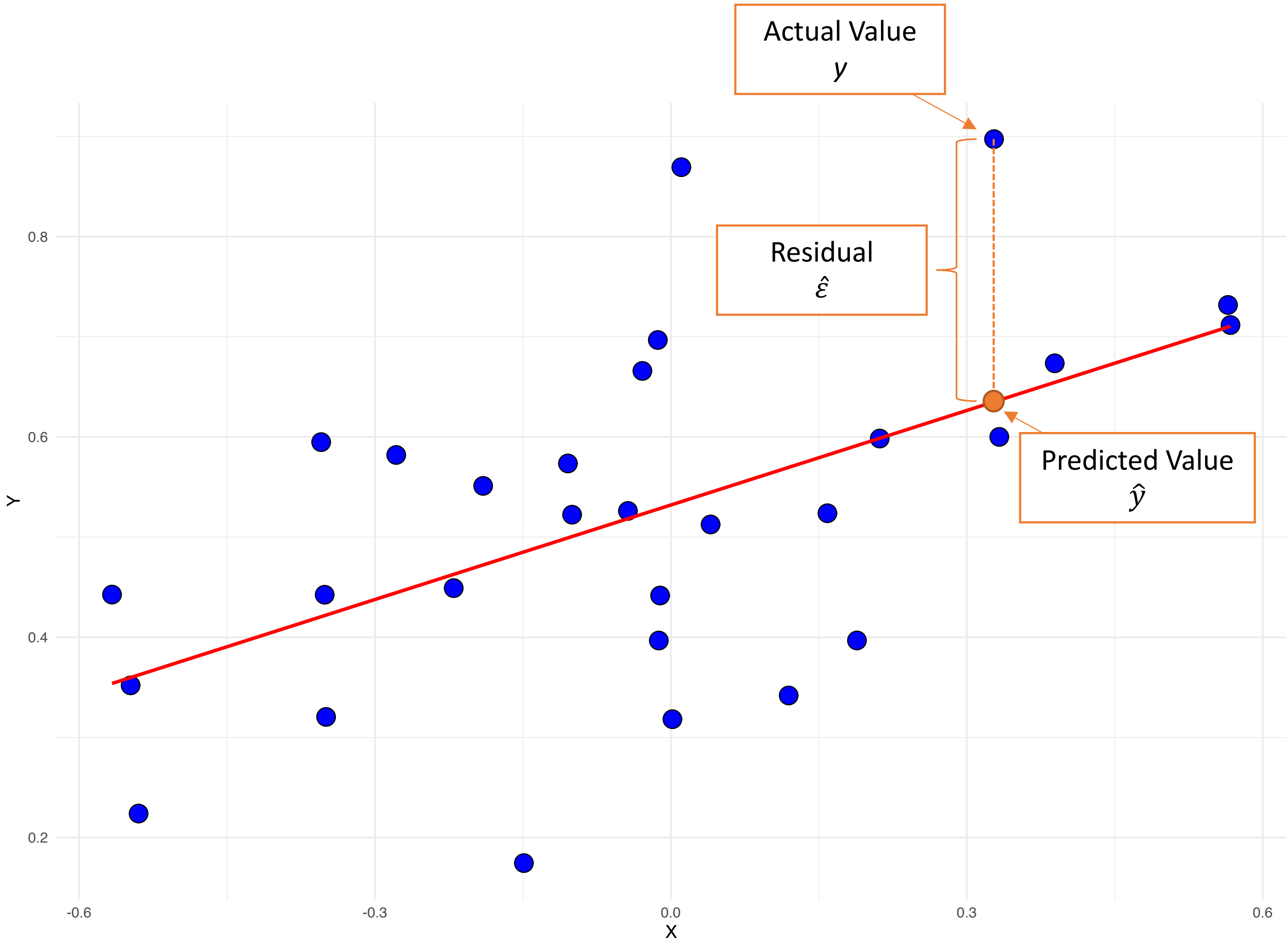
- Linear Regression is the first **statistical model** we will examine in this class. It differs from statistical tests because it is designed to **predict** values for observations we haven't seen yet.
- Specifically, this is a type of **linear model** – it's a method that draws a fitted line through the data, and uses that line to predict what the value of y will be for every value of x.



When you fit a line to a bivariate distribution, you get a new **predicted value** of *y* for every value of *x*. This predicted value is labelled $\hat{y}$.

The difference between *y* and $\hat{y}$ is called a **residual** and labelled $\hat{\varepsilon}$. The residuals are the changes in *y* that are not accounted for by the values of *x*.

The **line of best fit** is the line which has the smallest total residuals. The most common way to calculate this is with **ordinary least squares** (**OLS**), which adds the squares of all the residuals and finds the line with the smallest resulting value.



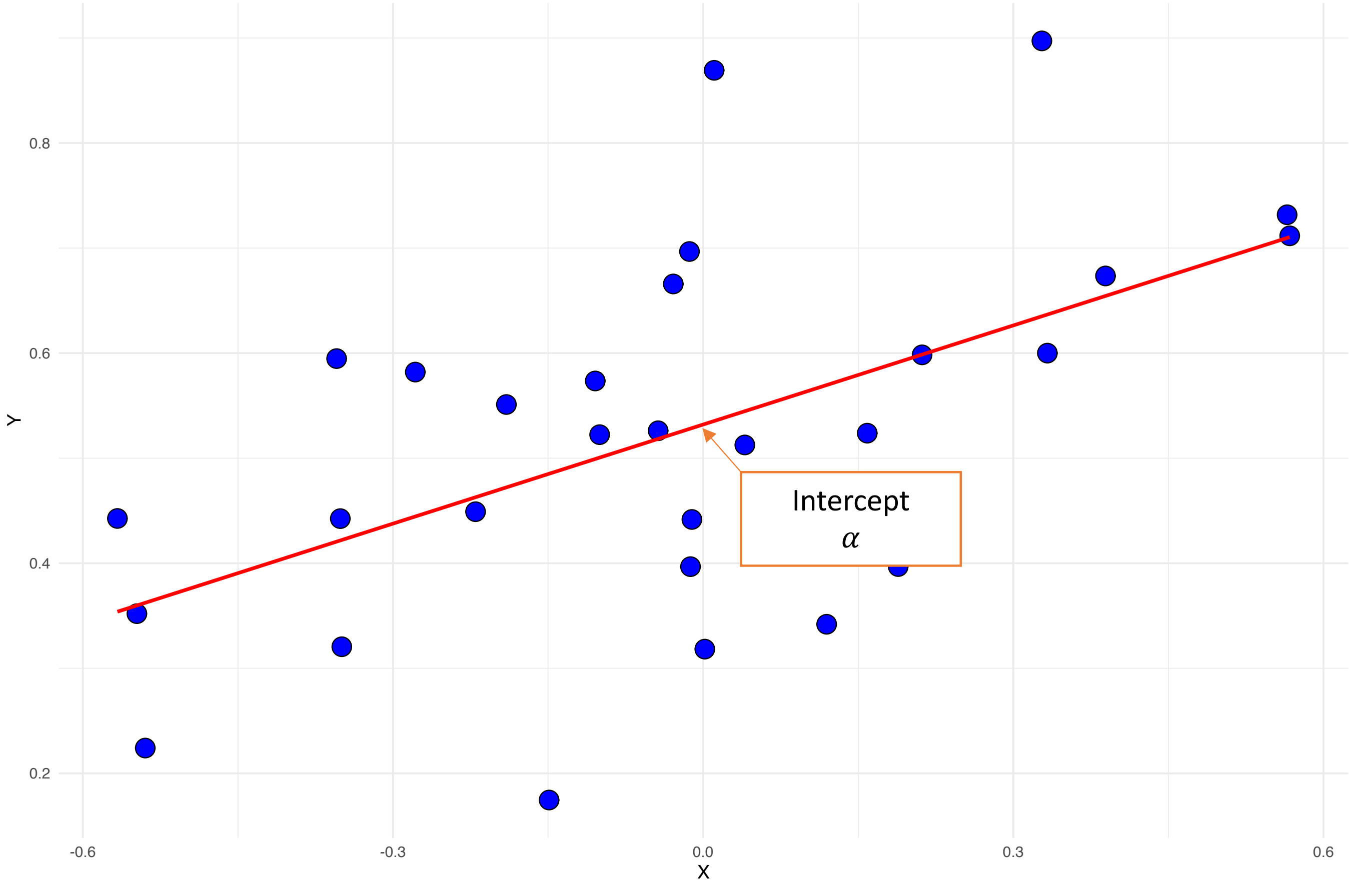
This **linear model** can be described by a simple equation:

$$y = \alpha + \beta x + \varepsilon$$

… meaning that *y* is predicted by a combination of *x* with two **coefficients**:

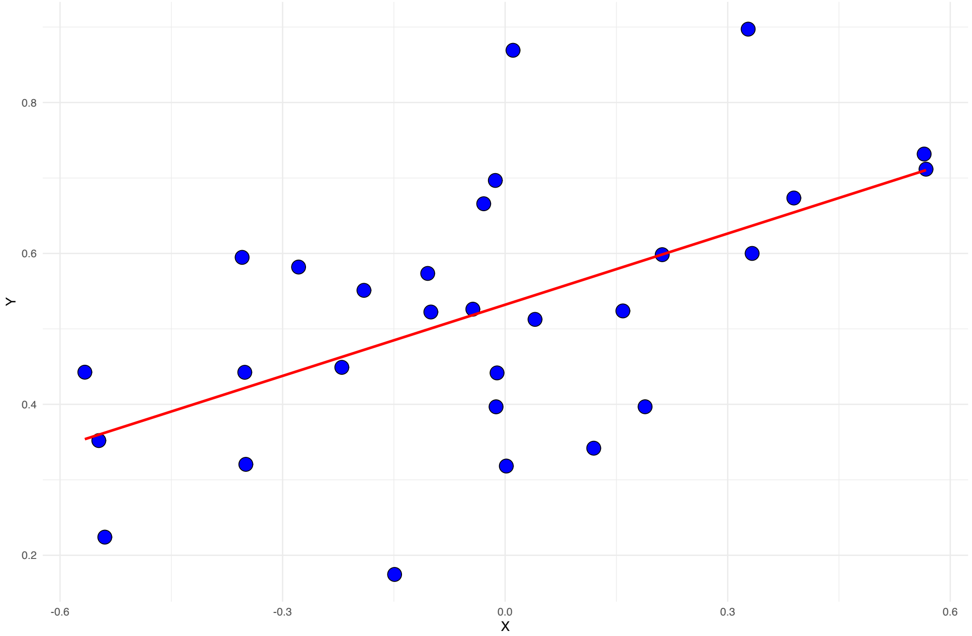
- The intercept α, which is the predicted outcome $\hat{y}$ when *x*=0;
- and the line slope β, which is the predicted increase in the outcome $\hat{y}$ for every one-unit increase in *x.*

Finally, the residuals we can't account for are included, and are called the error term, ε.



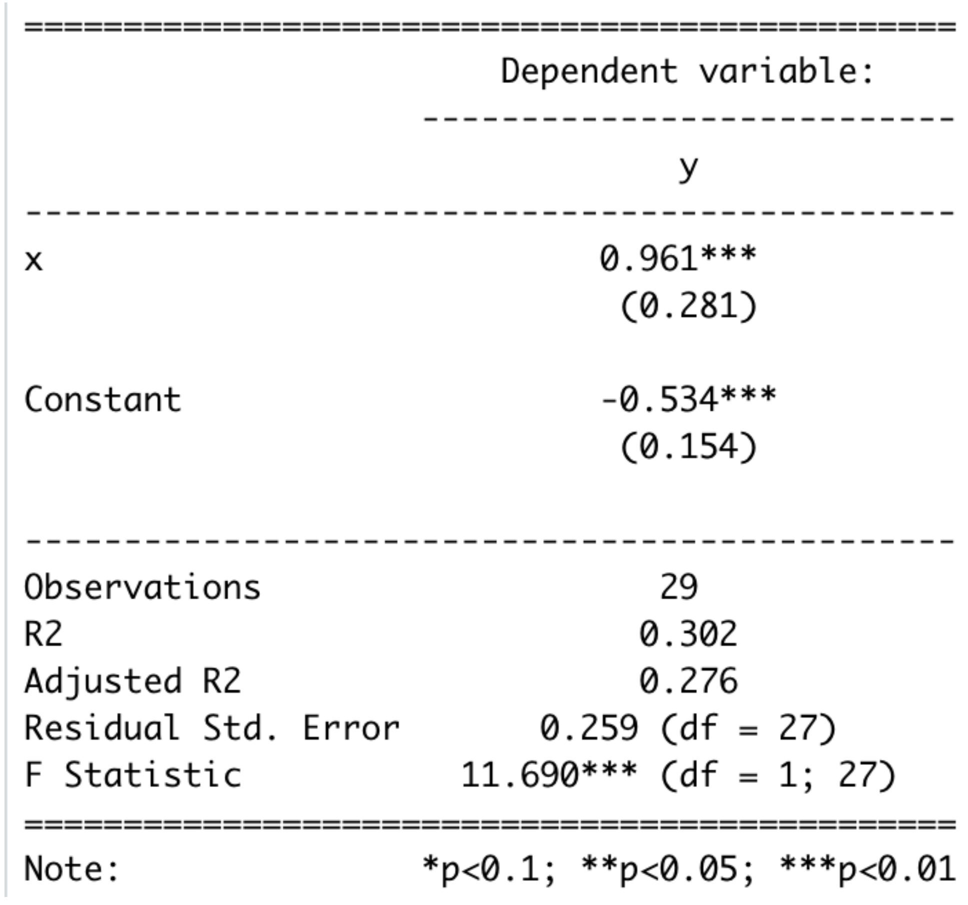
How well does the model fit?

- As with other tests we have done so far, OLS reports a ***p*-value** which we can interpret using a confidence level.
    - This is calculated by doing a T-test on the slope – the null hypothesis being that the slope is equal to zero.
- It also reports a value called **R^2**, or the **coefficient of determination**, which tells us what percentage of the change in *y* is being explained by *x*.
- In some fields, a high value of R^2 is expected, but when studying complex phenomenon with many inputs and factors (such as political or economic issues), a relatively low R^2 may be acceptable.



Presenting Linear Regression Results

- Linear regression results are commonly presented in a table like the one seen here.
- For each variable, including the intercept (here called the **Constant**), the **regression coefficient** and **standard deviation** are reported. Here, the p-value is indicated with stars; often it will be directly reported as a statistic.
- It's also common to report **R^2** and the **F Statistic**, which is a measurement of whether your model explains the data better than a model with no independent variables.
- There are variations on this table (it's common to report **95% confidence intervals**), but the key information reported remains the same.



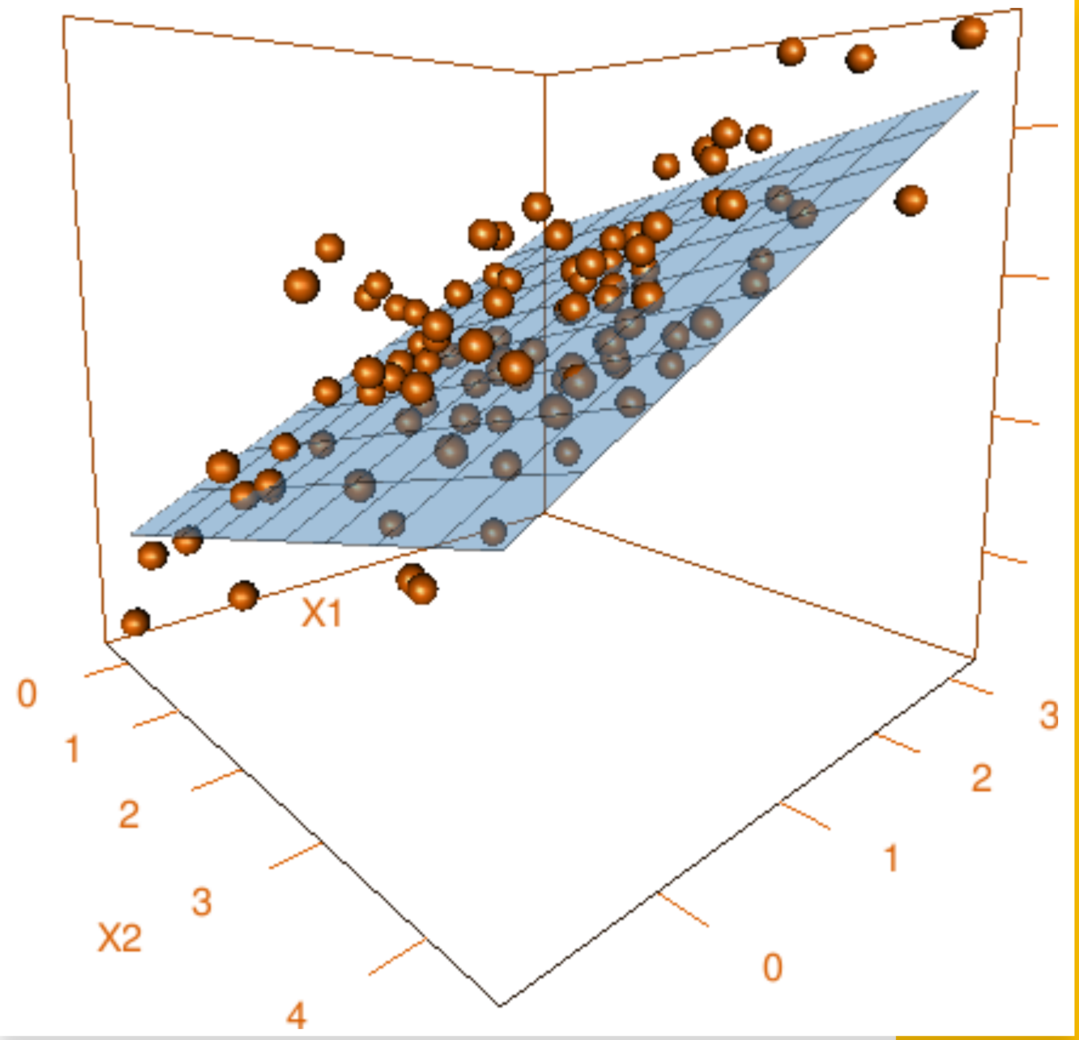
```
===============================================
                        Dependent variable:
                    ---------------------------
                                 y
-----------------------------------------------
x                            0.961***
                              (0.281)

Constant                     -0.534***
                              (0.154)

-----------------------------------------------
Observations                    29
R2                             0.302
Adjusted R2                    0.276
Residual Std. Error       0.259 (df = 27)
F Statistic           11.690*** (df = 1; 27)
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

Multiple Regression

- We've discussed regression with two variables – using variable *x* to explain outcome *y* – but often, we have more than one variable which we think are related to the outcome.
- We may include extra variables either because they are part of our hypothesis, **<u>or</u>** because we know they influence the outcome and wish to exclude their influence so we can see the effect of our main independent variable more clearly – these are called **<u>control variables</u>**.
- We can still use OLS for linear regression with multiple variables!
  - Adding variables adds **dimensions** to the model: with two predictor variables, the linear model is now finding a 2D plane, not a line.

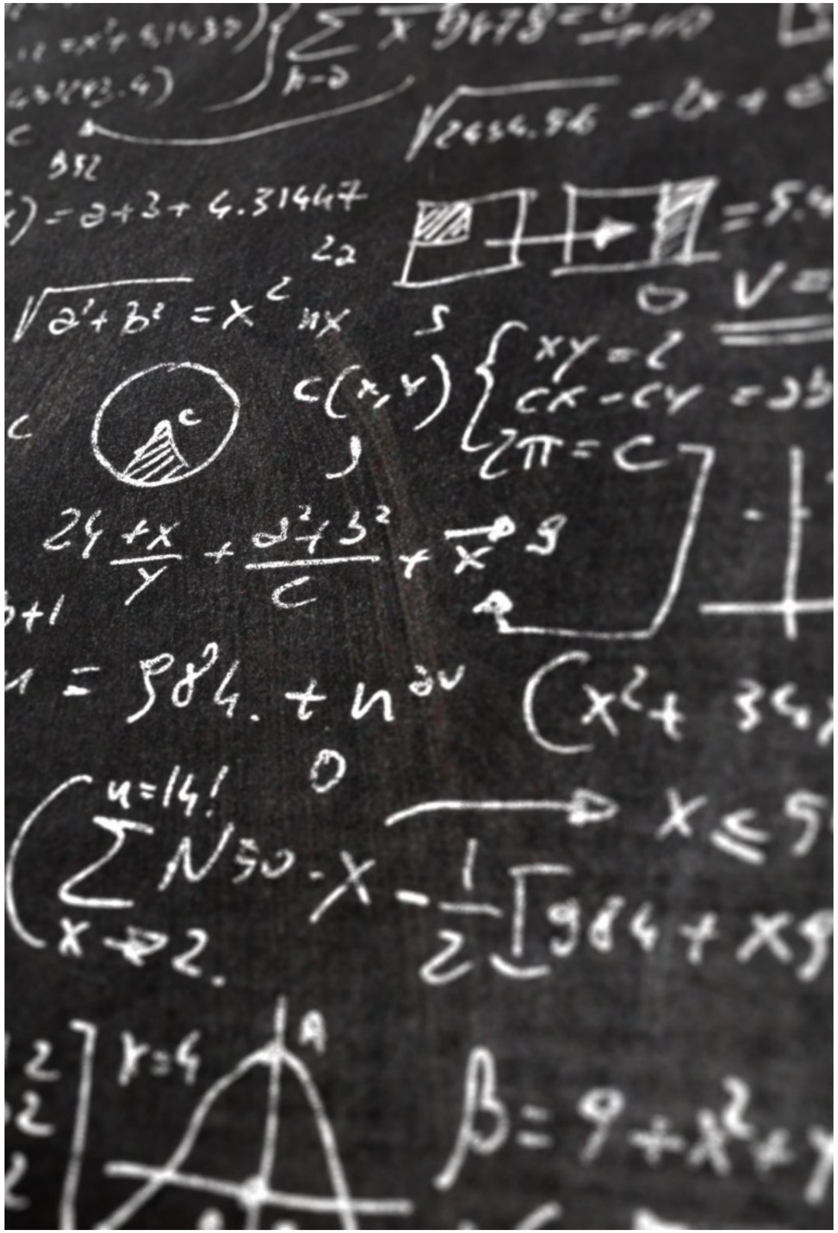


Multiple Regression Equations

- We express multiple regression models in equations in the same way that we expressed simple linear regression:

$$y = \alpha + \beta x_1 + \beta x_2 + \cdots + \beta x_n + \varepsilon$$

- $x_1, x_2, \ldots, x_n$ are all the **predictor / control variables** we wish to include (*n* being the number of variables).
- Just as before, α is the **intercept** and ε is the **error term** which contains the **residuals**.



Multiple Regression Assumptions

- While multiple regression is more complex mathematically than single regression, it’s no more difficult in R – you just specify as many extra variables as you want.
- All linear regressions assume that there is a linear relationship in your data – if the relationship is non-linear, you need to model it in a different way.
- However, multiple regression requires some extra assumptions about your data, which you should test to make sure you’re accurately modelling the relationship between variables.

Independent and Identically Distributed (IID) Errors

- IID (or i.i.d.) is a condition most sample data should meet. It’s made up of two sub-conditions:
- **Independence**: the value of each observation is not influenced in any way by the values of earlier observations. When observations influence each other, this is called **autocorrelation**.
- **Identical Distribution**: the observations should be scattered with the same degree of variance at all times. If the variance increases or decreases as more observations are made, this is called **heteroscedasticity**.
- In regression analysis, **the error terms (residuals) should be i.i.d.**
- If instead the errors show a specific pattern – autocorrelation or heteroscedasticity – this means **your model lacks an important explanatory variable.**

Minimal Multicollinearity

- Multicollinearity is a condition where two or more of your **independent variables are correlated with each other**.
- We're interested in knowing how X_1 and X_2 influence Y – but if X_1 and X_2 are also correlated with each other, it's impossible to tell which of them is responsible for the changes in Y.
- Some degree of multicollinearity is common. You might expect both income and education to influence vote choice, but income is probably also correlated with education.
- It becomes a problem at high levels (if two independent variables are perfectly correlated, OLS cannot fit a model at all and you'll get an error) and should be handled with care at low levels.
  - You can still fit a model that has some degree of multicollinearity, but the coefficient estimates for the correlated variables will be unreliable.

Adjusted R^2

- When we're using a single predictor variable, we can use a standard measurement of R^2 to show how much of the variation is being explained.
- However, with multiple variables we run into a problem – just adding extra variables to the model will usually increase R^2 a little bit, even if the extra variables aren't really helping at all.
  - This encourages people to "throw everything at the wall and see what sticks" – just adding loads of variables to the model in an attempt to get the best R^2 score. This is a form of ***p*-hacking** and should be avoided.
- A statistic called **Adjusted R^2** is usually reported by statistical software, and does a better job of handling extra variables; if a new variable isn't adding real value to the model, Adjusted R^2 will actually drop instead of rising.

Let’s move over to R